Otto-Friedrich-Universität Bamberg

Issues of corpus comparability and register variation in the International Corpus of English: Theories and computer applications

Fabian Vetter

Issues of corpus comparability and register variation in the International Corpus of English: Theories and computer applications

Fabian Vetter

Vetter, Fabian. 2021. *Issues of corpus comparability and register variation in the International Corpus of English: Theories and computer applications*. University of Bamberg dissertation.

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.d-nb.de/ abrufbar.

Diese Arbeit hat der Fakultät Geistes- und Kulturwissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Gutachter: Prof. Dr. Manfred Krug

Gutachter: Prof. Dr. Ole Schützler

Tag der mündlichen Prüfung: 10. Juli 2020

URN: urn:nbn:de:bvb:473-irb-524063

DOI: https://doi.org/10.20378/irb-52406

Open Science Framework (OSF) repository: https://osf.io/ztfsx/

Acknowledgments

This book grew out of a dissertation project at the Chair of English and Historical Linguistics at the University of Bamberg. I owe gratitude to many who, consciously or unconsciously, supported me along the way.

First and foremost, I would like to extend my gratitude to my supervisor, Manfred Krug, for finding the perfect balance between providing guidance and allowing me to follow my own path and to develop my own ideas. I also would like to thank Julia Schlüter for taking me under her wing and all the interesting projects we have worked on together; Ole Schützler and Valentin Werner for their critical and constructive feedback over the years; Gabriele Knappe and the colleagues of the *Linguistische Werkstatt* and *The Bamberg Graduate School of Linguistics* for sharing their fascination with a rich selection of linguistic topics; Lukas Sönning for sharing his passion for statistics; Catherine Irvine, Laura Murphy and Michaela Hilbert for proofreading this book; and all other colleagues not mentioned here with whom I had the pleasure of working with.

Finally, I would like to thank my friends, my family and Katharina, all of who not only endured me during all this time, but also supported and believed in me.

Contents

List of abbreviations

ACORN	A Classification of Residential Neighbourhoods
BNC	British National Corpus
CL	Corpus Linguistics
COCA	Corpus of Contemporary American English
CWB	Corpus Workbench
ESP	English for Specific Purposes
HCA	Hierarchical Cluster Analysis
ICE	International Corpus of English
IE	Institutional editorials
LLC	London Lund Corpus
LOB	Lancaster-Oslo/Bergen Corpus
MAT	Multidimensional Analysis Tagger
MDA	Multidimensional Analysis
MDS	Multidimensional Scaling
NB	Naïve Bayes
NJ	Neighbour-joining
OSF	Open Science Framework
PCA	Principal Component Analysis
PE	Personal editorials
POS	Parts-of-speech
SFL	Systemic Functional Linguistics
UWI	University of the West Indies
WIGUT	West Indies Group of University Teachers union
WWC	Wellington Corpus of Written New Zealand English

List of figures

List of tables

1. Introduction

Since the early beginnings of modern corpus linguistics (henceforth CL), which is often dated to the release of the Brown corpus in the early ‘60s, the use of corpora and corpus-linguistic methods have penetrated nearly all linguistic disciplines. The broad coverage of chapters in handbooks such as *The Routledge Handbook of English Corpus Linguistics* (O'Keeffe & McCarthy 2010), *The Cambridge Handbook of English Corpus Linguistics* (Biber & Reppen 2015) or *Corpus Linguistics: An international handbook* (Lüdeling & Kytö 2008, Lüdeling & Kytö 2009) vividly illustrate the wide range of methods and applications of corpora in the various linguistic disciplines.

While any collection of texts can theoretically be referred to as a corpus, most linguists agree that a corpus has to fulfil certain requirements in order to be called a linguistic corpus. The most common ones are that this collection of texts is of finite size, representative of a larger population of some sort, machine-readable and in some cases serves as a standard reference (cf. McEnery & Wilson 2004: 29–32). The property of finite size is necessary to guarantee transparency and reproducibility of the scientific process.[1] Machine-readability denotes that the collection of texts needs to be available in a digitized form that allows electronic storage and processing. While sometimes controversially debated in the context of CL (see section 2.2.2), collecting a sample which is representative of a larger population is a core concept of any empirical sampling process and many of the modern English reference corpora[2], such as the British National Corpus (BNC), members of the Brown corpus family or components of the International Corpus of English, thus act as a standard reference for the language variety under investigation. As reference corpora are not created with any specific research question in mind, it is unsurprising that ever since the beginnings of CL, the compilation of balanced reference corpora has played an important role.

Linguistic corpora nowadays come in all shapes and sizes, ranging from text collections that include a wide range of registers, to those that include only a specific register or mode, to highly specific databases that include, for example, only child language or learner English. While the number of available corpora is ever growing and it has become impossible to retain an overview, some authors attempt to list the most widely used ones (see Lee (2010) for corpora of the English language, or Xiao (2008) and Ostler (2008) for a broader overview).

The ways how corpora can be classified is equally diverse, but the following properties commonly serve as basis for classification: the range of included registers (reference or general corpora, such

[1] If corpora were continuously expanded, as is the case for monitor corpora, it would be impossible to reproduce results of studies which used previous versions of a corpus, unless the user interface allows to restrict queries to previous iterations of the corpus (as is the case for COCA, for instance).

[2] There appears to be no single established scheme for categorizing or referring to the various types of corpora available. This study will follow Tognini-Bonelli (2010) and refer to corpora which include a wide variety of registers with the aim of representativeness of a certain regional English variety (such as the BNC or the Brown corpus) as *reference corpus.*

as the BNC vs. specialized corpora, such as the Santa Barbara Corpus of Spoken American English), the mode (spoken vs. written corpora), the time period covered (synchronic vs. diachronic, modern vs. historical corpora) and the number of languages included (monolingual vs. multilingual or parallel corpora). Monitor corpora take a special role as they are expanded continuously to allow the tracking of recent changes in language (regularly relying on internet resources as input material). Due to their availability and broad range of applications, the most widely used English corpora are certainly synchronic reference corpora, such as the BNC or the Corpus of Contemporary American English (COCA). Collections of comparable corpora, such as the Brown corpus family or the components of the International Corpus of English (ICE), present a special case of synchronic reference corpora, as the individual corpora (or components) of these families were created with comparative studies in mind.

Although the creation of any corpus is already a demanding task, compiling a reference corpus is even more challenging – it should represent a microcosm of an entire language variety and as such must include a wide variety of situations where language is used. These distinct situations of language use are referred to as *registers* (cf. section 2.1.2). With representativeness as the goal, collecting data of different registers demands a great deal of planning and legwork from corpus compilers. Not all sampling techniques are applicable and sensible for all registers (cf. section 2.2.2). The creation of comparable corpora complicates this convoluted process even further as cultural differences as well as practical constraints must be taken into account. Corpus compilers are faced with the apparently unresolvable conflict of finding a compromise between representativeness and comparability (cf. section 2.2.3). On the one hand, components of a corpus family should be representative of the target varieties and, on the other hand, comparable to other components of the corpus family. To achieve comparability, the currently favoured course of action is to use the same sampling scheme for all components, that is, to sample the same number of texts from the same registers.

One aspect about this approach that could turn out to be problematic is that these registers are only defined fuzzily. In this context, *fuzzily* means that only register labels and very superficial descriptions are available – instead of operationalizing and delineating the individual registers through a window of acceptable configurations of text-external criteria. As this is not available for most corpora, the decision-making process informing the categorization of texts into registers remains — except for the written/spoken divide — largely opaque. And while frameworks for operationalizing register text-externally have been proposed — with Biber's (1988) or Biber & Conrad's (2019) *situational characteristics* certainly being the most influential in a CL context — relevant principles yet need to find their way into the compilation and annotation process. Indeed, most corpora come with only a limited amount of metadata, which also renders a reconstruction of the text-external criteria for some texts next to impossible.

For comparable corpus families, such as ICE or the original Brown quartet, where the same sampling scheme is used for multiple components, a non-operationalized or opaque definition of registers may still lead to small-scale differences between individual components. While the *press editorials* section, for instance, may contain only a very specific type of editorial in one member of a corpus family, it could very well also contain a range of types of opinion pieces in another. If the comparison for which these corpora are then used is sensitive to register variation, this may result in an over interpretation or, in the worst case, even to a misinterpretation of observed patterns. This issue might become even more problematic if various reference corpora are compared which are not part of a comparable corpus family but share the same or similar register labels.

To illustrate this problem further, let me return to the ICE section *press editorials*. As a subcategory of *persuasive writing,* texts in this section are defined as "distinguished from general news reports on the grounds that their main intention is to persuade rather than to inform. They are less directly tied to current events, and they afford the writer the opportunity to be discursive in a way that news journalism does not" (Nelson 1996: 33). This definition alone allows for a narrow and a broader interpretation of this category. In the narrow sense, it would contain only institutional editorials, i.e. opinion pieces that represent the point of view of the newspaper and are issued and edited by an editorial board. A broader interpretation would also include columns, comments (or *op-eds*), letters to the editor[3] as well as institutional editorials — in short, the full range of journalistic texts usually found in the editorials section of a newspaper. Although the aforementioned, non-institutional opinion pieces are sometimes dubbed *personal editorials,* they differ significantly from their institutional counterparts in some respects (cf. section 5.2.1). Where the Brown corpus family consequently follows the broader definition, some components of ICE only include *institutional editorials,* while others contain either only *personal editorials* or a mix of both (cf. section 3.1.1.2). Even though the various opinion pieces are technically distinguishable via text-external criteria (i.e. their situational characteristics), this kind of information is rarely included in the metadata.

The issue of shared text category labels containing texts from different registers becomes even more problematic when non-parallel corpora, such as the BNC or the COCA, are compared. Both contain sections sharing the same or very similar register labels (e.g. *newspaper* or *academic prose*), but these are divided into different subsections. While this difference is visible in the sampling scheme, the exact composition of the subsections remains opaque. And although the BNC contains a wealth of bibliographical details, evaluating the corpus comparability proves to be difficult as COCA contains only the bare minimum of metadata (as is the case for many other English corpora).

The majority of authors of corpus-based studies naturally exhibit a very reflected use of corpus resources and take register into account, yet these finer details of the sampling process (cf. section

[3] Although letters to the editor are strictly speaking not a journalistic text variety as they are composed by external, non-professional writers, they are still considered here, because some corpora (e.g. the Brown family as well as some ICE components) include them in the press editorials section.

2.2.2) are often disregarded. One of the main reasons for this is doubtlessly that this kind of information is either not available for many corpora or exceedingly hard to establish. Yet it is without a shadow of a doubt that additional information on corpus texts – be it details on the speakers or authors, such as age, gender or social class, or information on the target audience, topical domain or medium – empower linguists to reveal otherwise hidden patterns, to exploit the full potential such text collections offer and to evaluate the comparability of corpora.

In the absence of such information, this study consequently argues that compositional differences between corpora remain undiscovered (which in turn may obscure interesting linguistic patterns) and that these differences can reduce the comparability of corpora. Especially if comparable corpora (ICE used as an example) are used for comparative studies on linguistic phenomena which are sensitive to register variation, this study further argues that methods for detecting such undocumented register differences are required.

With the primary aim of uncovering undocumented compositional differences between components of ICE and presenting methods for detecting such, a blend of qualitative (mainly close reading and review of available metadata) and quantitative methods is employed. The quantitative methods are based on visualizing similarities and differences between profiles of corpus texts. These profiles comprise the frequencies of linguistic variables per text and will also be referred to as *frequency profiles*. Relevant theoretical foundations, the data and applied methodology are discussed in sections 2 and 3. Definitions of important terms related to text categories are provided in section 2.1, and the fundamentals of an empirical sampling process with respect to corpus design and corpus comparability are discussed in section 2.2. Section 2.3 reviews methods for clustering texts and section 2.4 reflects on the role of corpus linguist(ic)s in general. After a brief interim summary (section 2.5), the detailed scope and aims of this study are presented in section 2.6. The components of ICE included in the analysis, the sampling of additional material as well as the compilation of the frequency profiles are described in section 3.1. Utilized statistical methods are detailed in section 3.2.

Another major goal of this study is the development of an interactive computer program (*ICEtree*) that grants fellow linguists the opportunity to apply the methods used in this study to components of ICE and registers that are not investigated in this study. *ICEtree* and its functions are described in section 4. The source code and a detailed description of *ICEtree* are accessible in a repository in the Open Science Framework (OSF, https://osf.io/ztfsx/).

Section 5 then presents four case studies that utilize *ICEtree* to investigate compositional differences between various components of ICE from a register perspective. The first case study scrutinizes the *press editorials* section of the British (ICE-GB), Canadian (ICE-CAN), Jamaican (ICE-JA), Indian (ICE-IND), Hong Kong (ICE-HK) and American (ICE-USA) components of ICE. The second case study focuses on the section *skills and hobbies*, the third the section *business transactions* and the last the section *phone calls* in ICE-GB, ICE-CAN, ICE-JA and ICE-IND. In all of the sections investigated, it becomes apparent that the composition differs between components of ICE. The final section discusses the implications of the findings, points out potential avenues for future research and consequently advocates that an extended annotation of corpora could alleviate the issues addressed in this study.

2. Theoretical foundations

The case studies presented in this study as well as the application *ICEtree* itself draw on a range of (linguistic) concepts. This chapter will provide necessary terminological definitions and describe relevant concepts. The first section (2.1) describes possibilities of how text can be grouped with a focus on situationally defined text varieties and provides definitions of relevant terms. Section 2.2 outlines fundamental ideas of empirical sampling, reviews current practices in compiling linguistic corpora against this background and identifies potential problems when linguistic corpora are used for comparative studies. Section 2.3 sums up approaches to the quantitative description and clustering of texts. The next section (2.4) reflects upon the nature of the linguistic discipline corpus linguistics and sketches possible future developments. After an interim summary and a critical reflection (2.5), the final section (2.6) details the scope and aim of this study.

2.1. Text categories and situationally defined language varieties

Literate humans have the ability to intuitively group texts, regardless of whether they are spoken or written, into more or less clear-cut categories. And although we have little trouble grouping texts by various criteria, it seems that we perceive the situation in which language is used as most informative – it is, after all, the situation to which we naturally tailor the way we speak and write. The idea that language users shape their language to fit specific situations of use has a long history in linguistics and has been discussed in the context of various linguistic traditions. Sadly, this has also led to a terminological jungle and, much worse, to authors using the same term for different notions – this is especially true for *register* and *genre*. The purpose here is not to give a comprehensive review of the conceptualizations of individual authors (for this, see Lee 2001 or the respective chapters in Bawarshi & Reiff 2010 or Biber & Conrad 2019). Instead, the aim of this section is to provide definitions of relevant terms that will be used for the remainder of this study and outline the core ideas of relevant concepts. Yet, especially for the terms *register* and *genre*, some discussion seems unavoidable. By and large however, I will follow Biber & Conrad (2019) in their definition of *register* and avoid the term *genre* for the sake of clarity.

2.1.1. Text variety, text category and text type

The terms *text variety* (e.g. used by Biber & Conrad 2019) and (*text*) *category* (e.g. used by Johansson 1978 in the LOB corpus manual) are probably the most neutral terms for referring to a grouping of texts. These terms neither imply any theoretical connection nor any particular set of variables on which the grouping is based (e.g. situation of use). Where *text variety* simply acknowledges that texts vary, *text category* sets a stronger focus on grouping texts into distinct categories. Both terms can be used on any level of generalization as they also do not imply any kind of basic-level status of categories, which is also why Lee (2001: 49) dubs *text category* as a “catch-all term”. However, as it is often important why texts are part of a group, these terms may be too generic in many linguistic contexts.

In contrast, the definition of *text type* varies with no consensus among the various authors (again, see Lee 2001 for a discussion). While this term is sometimes used synonymously to *text category*, some authors (e.g. Biber 1988: 70, Diller 2002: 2) define *text types* as linguistically demarcated groupings of texts (i.e. based on text-internal features).

I will use the terms *text variety* or *text category* as theory-neutral terms for referring to groups of texts, regardless of the defining criteria, and *text type* for groupings of texts based on text-internal criteria.

2.1.2. Genre and register

As stated above, both terms, *genre* and *register*, have been used in the context of several linguistic traditions and by many authors. Lee (2001) attempts to disentangle the terminological patchwork surrounding these terms and reviews the use of *genre* and *register* as used by various authors. Instead of reiterating the conceptualization of individual authors, however, I will instead outline and compare the central ideas surrounding these terms from three influential traditions which exhibit the most developed ideas, namely Systemic Functional Linguistics (SFL), English for Specific Purposes (ESP) following Swales (1991) and the Multi-Dimensional Analysis (MDA) of registers as outlined in Biber & Conrad (2019). Afterwards, I will also discuss Lee's (2001) synthesis. At the outset of this section, it must be emphasized that the linguistic traditions addressed here do not only differ in the conceptualization and definition of the terms *genre* and *register*, but also in their fundamental goals. Where the goal of studies in an ESP context is to improve language teaching, SFL sees language as a social activity, seeks to explain how it is used for creating meaning and aims to provide a comprehensive framework to analyse language use from this perspective. Although certainly influenced by the systemic-functional view, the goal of the MDA framework is to explore and explain co-occurrence patterns of lexico-grammatical features present in situationally defined language varieties. Although these traditions and authors define the terms in question to some extent differently, there is also a significant overlap of ideas. At the end of this section, I will propose working definitions of these terms which will be used for the remainder of this study. As this study is deeply rooted in corpus linguistics, special attention will be given to the empirical collection of texts and the compilation of a linguistic corpus.

2.1.2.1. Register and genre in various linguistic traditions

A *register* is in both the SFL and the MDA framework essentially described as a situationally defined language variety and the occurrence of linguistic features, such as the use of nominalizations, is explained via text-external criteria or the situation of use. The ideas in SFL (cf. Eggins 1994: 49–80) are, however, more abstract compared to the more descriptive orientation of the corpus-based MDA framework. In SFL, the context of situation of a register is described in terms of three register variables *mode, tenor* and *field.* Field denotes the topic of the text, tenor the social relationship between the participants and mode the role language plays. The occurrence of linguistic features is reflected and explained with reference to these register variables. They are further linked to semantic planes and thus can also be analysed in terms of which kind of meaning they encode (cf. Eggins 1994: 11–24). While the exemplary register analyses provided by Eggins (1994: 54–57) illustrate the explanatory power and flexibility of the SFL framework, it also becomes apparent that its strengths lie in qualitative register analyses and that it was neither intended nor is it suited for classificatory purposes of text varieties on a larger scale.

In the MDA framework, the text-external criteria are referred to as *situational characteristics* and a more exhaustive set of variables is provided (cf. Biber & Conrad 2019: 40, Figure 1). This set includes for instance a description of the participants of the text (including gender, age and the social relationship between those), the mode and the production circumstances of the text. A range of possible nominal values for each variable is provided so that texts of many registers could be adequately annotated and further subdivided. Although linguistic features are also analysed from a functional perspective, the MDA framework lacks the social semiotic approach present in SFL. Instead, and also probably due to the strong quantitative and corpus-based orientation of the MDA framework, texts are analysed for co-occurrence patterns of surface level features, such as the frequency of special verb types (e.g. modals or private or public verbs) or syntactical structures (e.g. subordination and complementation patterns). The statistical method employed in the MDA framework (i.e. factor analysis) is used in more recent works with other feature sets as well (e.g. Grieve 2014), but many studies still rely on the original tagger, methodology and feature set as described in Biber (1988: 59–99) and Biber (1989). Other studies employ the Multidimensional Analysis Tagger (MAT) (e.g. Crosthwaite 2016, Kruger & Smith 2018, Montoro 2018), a freeware tool that accurately replicates the methodology of the original tagger (Nini 2019). Groups of co-occurring linguistic features are then interpreted in terms of broader communicative functions and situations they are associated with (i.e. dimensions of variation), e.g. the communication of abstract information or interactive, personal communication and the expression of feelings.

Authors working in the context of ESP predominantly use the term *genre* – the notion of register and the functional analysis of linguistic features is only later incorporated and is reminiscent of the MDA framework (e.g. Bhatia 2003). Forming a central concept in ESP, genre is defined by Swales (1991: 58) as follows:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community. The genre names inherited and produced by discourse

> communities and imported by others constitute valuable ethnographic communication, but typically need further validation.

Compared to the concept of register, where text-external criteria (be it tenor, mode and field or situational characteristics) are used as a basis for grouping, the key criterion for grouping communicative events into genres is the communicative purpose. Additionally, the idea of prototypicality and exemplars of texts[4] is introduced. Although exemplars of genres are defined text-externally by their communicative purpose, it is up to the members of the discourse community to evaluate the prototypicality of an exemplar based on mostly text-internal criteria (structure, style and content). By writing that genres are owned by expert discourse communities (cf. Swales 1991: 26), Swales also takes into account that not all individuals are familiar with or competent in all genres. Where the focus of a register analysis is on lexico-grammatical patterns associated with particular situations of language use, a genre analysis in the context of ESP typically investigates the structural and rhetorical make-up of texts in a so-called move analysis. In a move analysis, texts or parts of such are analysed for their rhetorical structure. For this purpose, texts are segmented into smaller sections based on individual communicative purposes these parts aim to achieve (see Kanoksilapatham 2007 for an introduction to move analysis). Additionally, content and style are analysed in genre analysis. As the ultimate goal of ESP is to help learners of English to become competent in (often specialized) genres, the purpose of genre analysis is to distil the core features of prototypical texts of a genre and incorporate these findings into the teaching of English for Specific Purposes. While the main interest lies in the structure of texts, phrases and structures which commonly appear in certain moves are also part of the analysis.

Although the focus is not on teaching, the core ideas on genre in SFL are similar to those of ESP. In SFL, genre is defined as "linguistically-achieved activity types recognized as meaningful (i.e. appropriate) in a given culture" (Eggins 1994: 35) or as a "staged, goal-oriented, purposeful activity in which speakers engage as members of our culture" (Martin 2001: 155). In both SFL and ESP, genre is defined via communicative purpose and thus seen as a highly schematically structured activity that is realized through language. Similar to a move analysis, the analysis of this schematic structure of a genre in SFL segments a text into units which fulfil a specific purpose. Further, genre is said to place restrictions on content and style in both frameworks. Genre in SFL, however, is not discussed in the context of discourse communities, but rather in the contexts of cultures. Different cultures may possess different genre potentials (cf. Eggins 1994: 34–37), so that certain genres may exist in one culture, but not in another. One of the consequences is again that not all individuals are competent in all genres. In SFL, genre is conceptualized as more generic than register. Genres can invoke various registers and the invoked registers also may differ culturally. In contrast, while Biber & Conrad (2019) often use the term *genre* alongside *register* and do not

[4] Although communicative events do not necessarily coincide with texts since a text can be split into individual communicative events, I will gloss over this fact as it is not directly relevant here.

locate either on different theoretical hierarchies, their conceptualization of *genre* bears a strong resemblance to that prevalent in ESP. They write that "[g]enres generally have simple names in a culture, [...] are governed by specific conventions, generally recognized by members of a culture, and so the genre itself is named within the culture" (Biber & Conrad 2019: 34). In contrast to *register*, the presence of a conventionalized structure of a text variety is thus a required factor for a text variety to be considered a *genre*. Another crucial difference to the conceptualizations of ESP and SFL is that Biber & Conrad (2019: 36) propose using the same set of variables (i.e. situational characteristics) for the description of *genres* as well as *registers* – including communicative purpose. With regards to a register analysis, Biber & Conrad (2019: 45–47) further point out that communicative purpose can be analysed on several levels of specificity, that the purpose can change in the progression of a single text and that texts can possess multiple purposes. In terms of methodology, they additionally distinguish between a register and a genre *perspective*. When analysing language from a genre perspective, entire texts are the analysed for their structure and conventionalized, but non-pervasive features. For the analysis of the rhetorical structure, the authors adopt Swales' move analysis approach (cf. Biber & Conrad 2019: 163–169). In a register perspective, on the other hand, where pervasive linguistic features are considered, the use of text excerpts is sufficient.

Based on his review, Lee (2001) concludes that the terms *genre* and *register* essentially denote two different perspectives on the same object and tries to reconcile the various linguistic traditions by defining the two terms as follows:

> *Register* is used when we view a text as language: as the instantiation of a conventionalised, functional configuration of language tied to certain broad societal situations, that is, variety according to use. Here, the point of view is somewhat static and uncritical: different situations "require" different configurations of language, each being "appropriate" to its task, being maximally "functionally adapted" to the immediate situational parameters of contextual use. *Genre* is used when we view the text as a member of a category: a culturally recognised artifact, a grouping of texts according to some conventionally recognised criteria, a grouping according to purposive goals, culturally defined. Here, the point of view is more dynamic and, as used by certain authors, incorporates a critical linguistic (ideological) perspective: Genres are categories established by consensus within a culture and hence subject to change as generic conventions are contested/challenged and revised, perceptibly or imperceptibly, over time. (Lee 2001: 46)

His definition of register captures both the essence of the SFL view with its distinction of tenor, mode and field (e.g. Eggins 1994: 49–80) and the definition offered by Biber & Conrad (2019). As opposed to register, Lee's (2001) definition of *genre* remains relatively vague. The focus is on texts being part of culturally negotiated, purposive or functional groupings. He adds that genres can invoke multiple registers (Lee 2001: 46) and thus mirrors largely the SFL approach where genre is situated above register and is defined via cultural context and purpose.

2.1.2.2. Towards a working definition of genre and register

While I can see the value in distinguishing between study designs that focus on pervasive, functionally motivated linguistic features (register perspective) and those that focus on the structure and conventions of entire texts (genre perspective) as suggested by Biber & Conrad (2019), I fail to see the practical value of a theoretical distinction between genre and register based on text-external criteria, as is the case in SFL. Communicative purpose alone can certainly not be the only factor why members of a culture or discourse community recognize texts as belonging to a certain group. Consider, for example, an introductory university lecture and a university text book for undergraduate students. Both share a very similar if not identical communicative purpose, yet they differ markedly in other register variables. Concerning the communicative purpose, both the lecture and the textbook aim to introduce a larger number of students to a very specific topic. Yet they differ markedly in their production circumstances, the mode and other register variables. Or consider newspapers: a collection of texts with vastly different purposes. Where some texts aim to inform the reader about current events (hard news), other texts try to persuade the reader of a certain point of view (e.g. press editorials). Readers of newspapers are certainly aware of the different purposes, yet it is hard to imagine that newspaper texts are more readily perceived to belong to a grouping based on communicative purpose (e.g. persuasive texts) rather than by medium (newspaper). Similarly, from a text-internal perspective, newspaper texts certainly form a more homogenous category than persuasive texts in general. When basing the categorization on only one variable, patterns associated with other text-external variables might be overlooked.

In any case, why should communicative purpose and culture (or discourse community) not be part of the notion of register, as implicitly proposed by Biber & Conrad (2019)? If register is a situationally defined language variety and we accept that texts are realizations of registers, does this then not already imply a culturally recognizable grouping of texts – namely by situation of use? What is to gain by separating communicative purpose and culture from other text-external variables? If language is investigated empirically – as is the case for any corpus-linguistic endeavour – the communicative purpose of a text and the cultural background in which it was created seem inseparable from other register variables in any case. Every form of verbal communication is characterized by a unique combination of text-external variables.

Largely following Biber & Conrad (2019: 6–11), I define a *register* then as follows:

i) A register is a language variety defined text-externally by its situational characteristics, including culture and communicative purpose.
ii) Consequently, a register can contain subregisters, and the exact composition of a register in terms of subregisters may vary culturally[5] and diachronically.

[5] This, of course, has ramifications for how regional variation can be defined and studied with corpus-linguistic methods (see discussion on corpus comparability in section 2.2.3).

iii) Registers can be clearly delineated by a window of acceptable configurations of situational characteristics and are not tied to a specific level of generality.
iv) Category membership is discrete, not gradual.

What follows from this definition is that register is an inherently multivariate phenomenon and, at least in a corpus-linguistic context, serves as means for structuring discourse. Multiple hierarchical schemes for categorizing texts with different defining criteria for different levels of specificity are conceivable. While it is established in corpus linguistics that the most common top-level distinction is based on mode (spoken vs. written), for example, other configurations are also conceivable. The emergence of digital registers, such as chats and text messages (e.g. WhatsApp or online forums), voice messages and video calls or online newspapers with a comment function blurs the previously relatively clear line between non-interactive written registers and predominantly interactive spoken registers. This definition of register, for example, would allow to distinguish texts based on interactivity first and on mode second.

The terms *register perspective* and *genre perspective* will be used as defined in Biber & Conrad (2019: 2): the register perspective takes a functional perspective and has its focus on pervasive linguistic features whereas the genre perspective focuses on conventionalized features and the structure of complete texts. Especially in a corpus-linguistic context, where we hardly ever have access to complete texts, we are often restricted to the register perspective.

It appears that much of the confusion surrounding the terms *register* and *genre* originates from the observation of prototype effects in the way we cognitively categorize texts.[6] Even if we include culture and communicative purpose in the concept of register, it is relatively easy to imagine that some texts of a register are *perceived* as more central, more prototypical than others. Think, for example, of a situation where you encounter a close friend in a professional context. Imagine further that said friend is an insurance agent and you want to procure some kind of insurance. The sales conversation you engage in will certainly be characterized by typical features of sales conversations, but it will likely also exhibit atypical traits due to the personal knowledge and the close, amicable relationship you share. The conversation may even switch between a sales conversation and a conversation with a friend on personal topics. While undoubtedly a sales conversation, we would hardly perceive this as a *typical* sales conversation. This conversation might even remind us rather more of a conversation with a friend who has expert knowledge in a certain field and is offering advice. If we were to analyse sales conversations empirically, we would likely try to find more typical examples or place more restrictions on the definition of a sales conversation. The type of interaction sketched here illustrates also that humans perceive the boundaries of such categories as fuzzy and that some texts exist in the space in between categories or are a mix of categories (sometimes also referred to as hybrids, hybrid genres or hybrid registers, e.g. Zhou 2012, Biber & Conrad 2019: 45–47).

[6] The reader unfamiliar with prototype theory is referred to Taylor (2011)

However, the definition of register offered earlier does not assume such an internal structure of categories. Instead, a register is defined as a hierarchically structured, clear-cut grouping of texts delineated by situational characteristics. And while registers may contain subregisters, category membership is discrete, i.e. texts either belong to one register or they belong to another. In short, the classical or Aristotelean way of categorization[7] is employed for register. In contrast, the definition of genre in ESP, SFL or in Lee's synthesis, assumes a categorization around prototypes. These two theories of categorization are antithetical in many ways, and both are used in many linguistic contexts (see Taylor 1995 or Aarts 2006 for an overview). Where the prototype theory draws fuzzy boundaries around categories and acknowledges that some members of a category are more central or prototypical (exemplars) than others, the classical theory of categorization draws clear-cut boundaries and all members of a category are of equal status. The role and status of properties that define category membership also differ in both theories. Where all defining properties must be present and are of equal importance for classification in the classical approach, in a prototypical approach, not all properties need to be present for an individual to belong to a category and some properties may be more relevant for determining category membership.

The notion that both theories of categorization are employed in linguistics is also voiced by Taylor (1995: 68–80) and by Diller (2002), albeit with a slightly different focus. Both authors distinguish between a prototype-based text categorization, the categories of which are referred to as *folk* or *natural categories*, and categories using the classical theory of categorization, which are dubbed *expert categories*. Diller (2002) sets his focus on the categorization of texts and describes *genres* as *folk* and *text types* as *expert categories*. However, he defines genres as fuzzy categories with a prototypical structure and mostly demarcated via text-external criteria, and text types as clear-cut categories defined via text-internal criteria. The terms *genre* and *register* as discussed in this study are both defined text-externally. Taylor (1995) discusses the use of prototype theory in linguistics in general and adds that "[f]olk categories [...] are grounded in the way people normally perceive and interact with the things in their environment" (1995: 72). With regards to expert categories, he writes that "one of the main activities of experts [...] is precisely the 'drawing of boundaries' (cf. Wittgenstein 1978: 33) around essentially fuzzy categories" (ibid.). He further argues that both categorizations often even coexist within each person. As a prime example for this apparent conflict, Taylor (1995: 68–70) discusses the categories odd number and even number. Although these are certainly expert categories on the one hand with clearly defined boundaries in which all members are of the same status, participants, when asked to rate the degree of membership of numbers to these two categories, still assigned a higher degree of membership to the numbers 2 and 4 than to even numbers above 4 for the category of even numbers. Although I prefer the term *natural categories* over *folk categories*, I generally concur with the assessment that natural categories appear to be more important in everyday usage, whereas expert categories, in the case of text varieties, represent the attempt by linguists to impose a hierarchical structure on discourse and

[7] An outline of the classical theory of categorization and how it is used in some linguistic domains is presented in Taylor (1995: 21–38).

that both types of categories serve different purposes. Lee (2001: 48–52), following up on Steen (1999), as well as Diller (2002) further observe that genres are not only based on a prototype based categorization, but may also present so-called basic-level categories, i.e. "categories [...] which are in the middle of a hierarchy of terms [...] and are characterized as having the maximal clustering of humanly-relevant properties (attributes), and are thus distinguishable from superordinate and subordinate terms" (Lee 2001: 48).

The definition of *genre* I propose here is very much similar to that of a register in that it is a situationally defined language variety, but that instead of the classical theory of categorization, a prototype approach is employed. The key differences therefore concern the way these categories are structured (i.e. their internal structure, their boundaries and possible hierarchies) and that human cognition is taken into account. The presence of an internal rhetorical structure as described in ESP or in Biber & Conrad (2019: 36) is not a required criterion. A genre is further said to be a basic-level category of text varieties, a grouping of texts that is cognitively perceived by human beings as relevant and practical in daily life. A hierarchy of superordinate and subordinate genres always takes the basic-level as the outset. As with register, the criteria delineating genres are text-external and the same set of variables can be used (e.g. those described in section 2.1.2.3). However, individual criteria may differ in their cue validity, so that the criteria that are most relevant for deciding membership may differ from genre to genre. Texts of a genre can be central or peripheral members of the category, and the boundaries of genres can be fuzzy. An overview of the main differences between the terms discussed in this section are summarized in Table 1.

The consequence of this definition is naturally that if it is to be used for data collection, a reliable way to determine the prototypicality of texts and the basic-level status of genres must be established. As these are cognitive phenomena, this can theoretically only be done by an individual who is part of the discourse community or culture in which the genre is used. In contrast, the concept of register allows the regrouping of texts based on situational variables that are relevant for the respective study. This is a crucial characteristic of the concept of register, as it facilitates the use of statistical modelling to investigate the effect certain situational criteria might have on a linguistic feature.

This should by no means be interpreted as downplaying the importance of a prototype approach to the categorization of texts – quite the contrary, in fact. Besides investigating the way we intuitively categorize texts, the study of prototype effects in all linguistic domains is certainly valuable and highly enlightening. The point I wish to make here is also not to establish that one theory of categorization is superior to or more important than the other. Instead, I would like to emphasize that the type of categorization used is linked to the research question, the way data are collected and analysed. Regardless of whether a genre or a register perspective is assumed, if we wish to investigate the characteristics prototypical texts of a category exhibit[8], a large-scale random collection of texts would probably be an ill-suited sampling method. However, if the goal is to describe the full range of variation a language user might encounter when consuming texts of a certain category, sampling only prototypical texts might result in a highly skewed picture. In addition, the sampling technique employed also affects which statistical methods can be employed.

Especially in a corpus-linguistic context, where data collection and analysis of the data are usually performed by a great number of different researchers, it is imperative to transparently document how the data are selected and how categories are conceptualized. Is the sampling based on a prototypical understanding of text categories (i.e. genres)? If so, how was the prototypicality of the individual texts and the basic-level status of categories established? Or does the classical theory of categorization form the basis for the definition of categories (i.e. register)? What were the windows of acceptable situational characteristics for the individual registers? How are the hierarchies defined? In general, however, I consider the concept of register better suited for corpus-linguistic investigations, not least due to its greater transparency (e.g. no fuzzy boundaries and equal status of category members), higher flexibility and quantifiability.

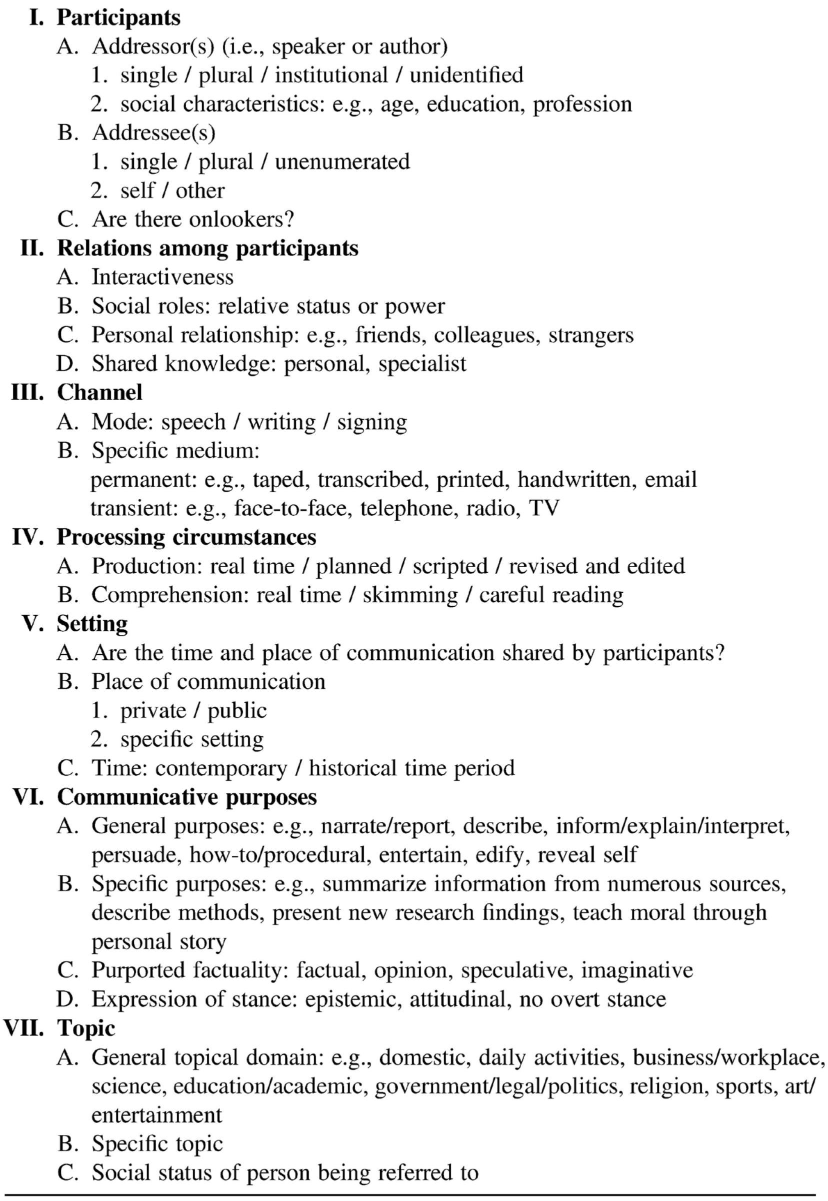
	Register	**Genre**	**Text type**	**Text variety, text category**
Definition	Text-externally defined language variety linked to specific situations of use	Text-externally defined language variety linked to specific situations of use	Coherent grouping of texts, defined text-internally	Neutral term to refer to groupings of texts
Theory of classification	Classical or Aristotelian theory	Prototype theory	Both may be applied	Both may be applied
Defining criteria	Situational characteristics	Situational characteristics + cognitive factors	Any number of linguistic features (e.g. *n*-grams, word frequencies, etc.)	NA
Implied level of hierarchy	Flexible	Fixed: Basic-level, but super-/subordinate levels may be defined	Flexible	NA

Table 1: Main differences between the terms *register, genre, text type, text variety* and *text category*.

[8] Of course, this assumes that a method for unambiguously establishing the prototypicality of texts of a category is available.

2.1.2.3. Register variables

Concerning register variables (i.e. situational variables for categorizing texts), some attempts have been made to compile lists, with Atkins et al. (1992), the Expert Advisory Group on Language Engineering Standards (Expert Advisory Group on Language Engineering Standards 1996) and Biber & Conrad (2019: 31–48) offering the most practical and comprehensive compilations. In contrast to Atkins et al. (1992), the last two organize the criteria into larger, meaningful groups. The text-external criteria proposed in Expert Advisory Group on Language Engineering Standards (1996: 11–15) are grouped into variables "concerning the origin of the text that are thought to affect its structure or content" (1996: 11), variables "concerning the appearance of the text, its layout and relation to non-textual matter" (ibid.) and variables "concerning the reason for making the text and the intended effect it is expected to have" (ibid.). Biber & Conrad (2019: 40) group the text-external criteria (or situational characteristics, see Figure 1) by variables relating to the participants, the relationship between participants, the channel, the processing circumstances, the setting, the communicative purpose and the topic. These criteria greatly overlap with those proposed by Atkins et al. (1992: 15–18). What they all have in common is that they are based largely, but not exclusively, on text-external criteria, such as participants, relationship between the participants, medium, production circumstances and topic. This apparent conflict is addressed in EAGLES (1996), where text-external criteria are split into *circumstantial* and *reflexive*. *Circumstantial* criteria are truly text-external criteria, that is, criteria that can be gathered without reading the text, such as mode, medium or author. *Reflexive* criteria, on the other hand, form an in-between category as they can be deduced from cues within the text – either how the text refers to itself or through interpretation of the reader (Expert Advisory Group on Language Engineering Standards 1996: 8). Communicative purpose, target audience, topic and style are examples of reflexive text-external criteria. Biber & Conrad (2019) do not address this problem directly, but instead use *situational characteristics* when speaking of both, circumstantial and reflexive text-external criteria.



I. Participants
A. Addressor(s) (i.e., speaker or author)
1. single / plural / institutional / unidentified
2. social characteristics: e.g., age, education, profession
B. Addressee(s)
1. single / plural / unenumerated
2. self / other
C. Are there onlookers?

II. Relations among participants
A. Interactiveness
B. Social roles: relative status or power
C. Personal relationship: e.g., friends, colleagues, strangers
D. Shared knowledge: personal, specialist

III. Channel
A. Mode: speech / writing / signing
B. Specific medium:
permanent: e.g., taped, transcribed, printed, handwritten, email
transient: e.g., face-to-face, telephone, radio, TV

IV. Processing circumstances
A. Production: real time / planned / scripted / revised and edited
B. Comprehension: real time / skimming / careful reading

V. Setting
A. Are the time and place of communication shared by participants?
B. Place of communication
1. private / public
2. specific setting
C. Time: contemporary / historical time period

VI. Communicative purposes
A. General purposes: e.g., narrate/report, describe, inform/explain/interpret, persuade, how-to/procedural, entertain, edify, reveal self
B. Specific purposes: e.g., summarize information from numerous sources, describe methods, present new research findings, teach moral through personal story
C. Purported factuality: factual, opinion, speculative, imaginative
D. Expression of stance: epistemic, attitudinal, no overt stance

VII. Topic
A. General topical domain: e.g., domestic, daily activities, business/workplace, science, education/academic, government/legal/politics, religion, sports, art/entertainment
B. Specific topic
C. Social status of person being referred to

Figure 1: Exemplary list of text-external criteria: Situational characteristics (Biber & Conrad 2019: 40). © Douglas Biber and Susan Conrad 2009, 2019. Reproduced with permission of Cambridge University Press through PLSclear.

2.1.2.4. Register variation and the linguistic variable

Even though the studies of Douglas Biber[9] (e.g. 1988) and the framework and methodology for analysing registers as presented in Biber & Conrad (2019) are highly influential and widely recognized in the linguistic community – a fact also emphasized by publication titles such as *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (Berber Sardinha & Pinto 2014) – a wide range of corpus-based studies disregard the multivariate nature of registers and treat

[9] Although not as frequently referred to in the English corpus-linguistic community, Koch & Oesterreicher(1985, 2012) also underline the fact that situationally defined language varieties essentially shape the language we produce.

register (or better: register labels) as a single variable to be controlled for in the analysis of linguistic phenomena. Biber et al. (1999) demonstrate the effect of register on a vast number of linguistic structures in the *Longman grammar of spoken and written English,* and Biber (2012) further cautions against ignoring register in lexico-grammatical studies. However, it is also necessary to recognize that the effect register has on a linguistic variable varies considerably: where *that*-omission appears to be strongly susceptible to differences in register (e.g. Biber et al. 1999: 680), Wolk et al. (2013: 404) report that register accounts only for a small part of the variation encountered in their investigation of the dative alternation. Similarly, Werner (2014) finds only minor register effects in his analysis of the present perfect in World Englishes. From a more methodological perspective, Gries (2015) argues for the use of more statistical models that factor in the hierarchically nested structure of corpora and finds register effects for particle placement. This list of studies could certainly be continued, but the point I would like to address in this context is the importance of such findings in the light of using comparable corpus families, such as ICE.

It is argued and exemplified throughout this study that some registers in ICE exhibit register variation between national components and that this reduces the cross-corpus comparability, and as a result the reliability of studies using ICE for studies of regional variation. At the same time, I must point out that this negative effect might be negligible if the variable under investigation is only affected to a minor degree by differences in register. In such cases, using ICE and disregarding these sometimes unavoidable differences is certainly warranted. This line of reasoning is in agreement with Greenbaum (1996: 5), who argues that as the ICE project progresses and small diachronic discrepancies become unavoidable, “the corpora will be sufficiently similar to justify global comparisons”. But this also means that it has to be determined – ideally a priori – how strong the effect of register is on the linguistic variable to be investigated and if the ICE components used exhibit such discrepancies in sampling.

2.2. Corpora as samples of populations

Despite the technological progress we have seen in the past few decades, the creation of a linguistic corpus is still no trivial endeavour. And how can it be if a corpus is meant to be “[...] a manageably small scale model of the linguistic material which the corpus builders wish to study” (Atkins et al. 1992: 17)? In the case of most corpora, this means that said small scale model should be representative of an entire language or language variety. Besides critical non-linguistic issues, such as funding or copyright, handbooks for corpus linguists single out three key considerations when creating a reference corpus (e.g. Tognini-Bonelli 2001: 55–62, McEnery & Wilson 2004: 77–81, McEnery et al. 2006: 13–21, Hunston 2008: 160–166, Clancy 2010): *representativeness, balance* and *size*, the details of which will be discussed in the course of this chapter.

Although certain differences in the exact definition of the concept of representativeness exist (McEnery et al. 2006: 13–16), the most common interpretation is that the material in a corpus is only a sample of a larger population which it should represent. In contrast to specialized corpora, which contain only a very restricted selection of registers, reference corpora (such as the BNC, the components of ICE or the Brown family) aim to be representative of a national variety of English at a given point in time. The BNC, for example, was designed as a model of British English of the 1990s (cf. Aston & Burnard 1998: 29). The components of ICE aim to be representative of the respective national variety at a given time (cf. Greenbaum 1996), and members of the Brown corpus family representative of the written language of the respective variety (British and American English) and time (1960s and 1990s).

While it is relatively easy to state what a corpus should be representative of in broad terms, achieving representativeness can be notoriously difficult and involves a series of complex methodological and practical decisions. As a general rule, however, the representativeness of any empirical sample hinges on *how* the sample was drawn from the population. Generally, randomized sampling techniques result in more representative samples than, say, convenience samples where a researcher selects samples solely based on their convenient availability. Although a wealth of publications is concerned with the drawing of (representative) samples, the sample design of corpora is rarely taken into account in corpus-linguistic studies, and only a few authors address the sampling of corpora and connected issues directly (Clear 1992, Haan 1992, Meyer 2004b: 40–45, McEnery et al. 2006: 13–22, Nelson 2010, Wattam 2015: 19–36, Woods et al. 1986: 48–58). In some linguistic disciplines, such as sociolinguistics, for example, sampling methods seem to have received more attention (cf. Buchstaller & Khattab 2013: 85–90). Similarly, guides to quantitative linguistics mostly cover the fundamentals of the sampling process (e.g. Brons-Albert & Marx 2010, Rasinger 2008: 45–52), but make no reference to the unique situation CL finds itself in (see the following sections). Publications on corpus linguistics or statistics in linguistics (e.g. Baayen 2008, Gries 2013, Levshina 2015, Weisser 2016) in general tend to set their focus rather on methods such as collocation analysis, cluster approaches or the use of statistical modelling in linguistic investigations even though sampling may have a direct bearing on the approaches and statistical techniques that are permissible.

In order to describe the special case for corpus compilers, it is important to first outline some fundamental concepts of statistical sampling. In essence, sampling theory is concerned with the intricacies of drawing a sample from a population, including the questions of "how data are selected, out of all the possibilities that might have been observed, whether the selection process has been under the control of investigators or has been determined by nature or happenstance, and how to use such data to make inferences about the larger population of interest" (Thompson 2012: 2). In what follows, I will briefly summarize the central concepts of sampling as described in reference works on sampling (e.g. Fuller 2009, Levy & Lemeshow 2008, Särndal et al. 2003, Thompson 2012). This overview is by no means meant to be exhaustive as I concentrate on the concepts relevant to the subsequent discussion of sampling in corpus linguistics.

2.2.1. Sampling – an overview

The fundamental steps in sampling include the definition of the population of interest (also known as the *target population*), the elements that make up the population (*elements*), the units for sampling (*sample units*), the list from which those units are sampled (*sampling frame*), the method for choosing the units (*sample design*) and the variables one wishes to study (*population parameters*). In surveys for example, the sample units are often individual persons or households. Sample designs can broadly be categorized into probability and non-probability designs. In contrast to non-probability designs, the selection of sample units in probability designs is performed randomly so as not to induce any (human) bias in the selection of sample units and to ensure that each element of the population has a known and non-zero inclusion probability. However, probability sample designs require the construction of a *sampling frame* for the selection of samples – in its simplest form, this could be a computerized list where all the elements are listed and from which we can access and select the sample units. In this instance, Särndal et al. (2003: 10–13) speak of *direct element sampling*. For cases where it may be difficult to obtain a list of elements for constructing a sampling frame, the authors add that the sample units may also consist of clusters or groups of elements – as is often the case in survey studies where the sampling frame consists of households or city blocks instead of individuals (*area sampling*).

Sampling can also occur at multiple stages with each stage having its own sampling frame and design (cf. Levy & Lemeshow 2008: 19–20). Särndal et al. (2003: 10–11) provide a list of requirements for a sampling frame and stress that "[i]t is important to construct a good sampling frame", but also point out that "[t]here is no sense in fixing a target population for which a good [sampling] frame cannot realistically be obtained within budget restrictions" (Särndal et al. 2003: 14). In this context, Särndal et al. (2003: 13–14) distinguish further between *target population* and *frame population*, where the former refers to the entire population of interest and the latter to the population as listed on the sampling frame. The authors further state that ideally, although in reality highly unlikely, these two are identical. Särndal et al. (2003) refer to the case where the target population includes more elements than the frame population as *undercoverage*, and use *overcoverage* for the reverse case (e.g. if the sampling frame includes deceased participants). If identical elements occur multiple times on the sampling frame, the authors refer to this as *duplicate listings*. The following description of the individual non-probability and probability sample designs is based on the respective chapters in Daniel (2012).

Within probability sample designs, Daniel (2012: 125–174) distinguishes the following major types[10]: simple random sampling, stratified sampling, systematic sampling and cluster sampling. Simple random sampling, where the sample units are selected randomly from the sampling frame, is probably the simplest probability design. Systematic sampling is similar to random sampling in that the first element is selected randomly from the sampling frame. All subsequent elements are then selected in fixed intervals (e.g. every *n*th element). In a stratified sampling process, the target population is divided into homogenous, non-overlapping subgroups and a random sample is taken from each subgroup. The sampling frame should take into account the

[10] More technical descriptions of the sample designs can be found in the respective chapters of Levy & Lemeshow (2008) and Särndal et al. (2003).

variables on which the stratification is based. Depending on whether the proportions of the strata in the population are reflected in the sample, stratified sample designs can be further divided into proportionate and disproportionate sample designs. Cluster sampling may be used if the target population is scattered widely or the construction of a sampling frame is impractical. In cluster sampling, not individual elements, but naturally occurring clusters of elements are selected randomly. Heterogeneity within the clusters and homogeneity between the clusters is vital for a good cluster design.

Non-probability designs, on the other hand, are used in situations where probability sampling is impractical or impossible (e.g. due to constraints in human resources or funding or when a sampling frame cannot be constructed). Four types of non-probability sampling are distinguished and described by Daniel (2012: 81–124): availability sampling, purposive sampling, quota sampling and respondent-assisted sampling. In availability sampling, population elements are selected because they are (conveniently) available. In purposive sampling, elements are selected based on inclusion/exclusion criteria and whether they fit the study objective. In terms of selection criteria, Daniel (2012: 88–91) lists central tendency (e.g. the selection of prototypical or extreme elements), variability (e.g. maxi-/minimizing the variability of the sample), theory/model development (e.g. sampling elements to dis-/confirm a hypothesis) and judgement/reputation. Quota sampling (also "purposive quota sampling" or "stratified purposive sampling") is similar to stratified sampling in that the population is divided into mutually exclusive subgroups, but the elements for the subgroups are not selected randomly, but until a certain quota is fulfilled for each subgroup using availability sampling methods. As with stratified sampling, quota sampling can be further divided into proportional and non-proportional quota sampling.

Daniel (2012: 66–80) lists the advantages and disadvantages of both probability and non-probability sample designs and considers the lack of the possibility to determine the sampling error as one of the greatest weaknesses of non-probability sample designs. Among other things, he sees non-probability designs as the better choice if very specific elements of the population are targeted or if the study goal is exploratory or to provide an illustrative example. Other authors are more restrictive in their assessment and state that we cannot extrapolate from findings gleaned from non-probability samples as it is impossible to determine how well the population is represented by the sample (e.g. Särndal et al. 2003: 529–530). Yet, these sample designs are widely used, if not for the simple reason that constructing a sampling frame is often costly and sometimes even practically impossible. As for probability samples, Daniel (2012: 66–80) considers these as the suitable choice for quantitative research designs, if statistical inferences are to be made from the data and if the sample should be representative for the target population.

2.2.2. Corpus compilation and current practice

In what follows, I will try to apply the terminology used in statistical sampling to the compilation principles of extant corpora. This is done to highlight issues that arise due to the very specific nature of CL and to critically outline current practices of corpus compilation pertaining to widely used reference corpora. The focus here shall be on the BNC, as it is arguably one of the most carefully compiled, best documented and widely used corpora; the Brown and the LOB corpus, as it is the first set of comparable corpora; and the ICE family, as it is also a widely used set of comparable corpora. To my knowledge, similar attempts made previously (e.g. Clear 1992, Meyer 2004b: 40–44, McEnery et al. 2006: 13–21) have not strived for the level of detail this section is aiming at.

2.2.2.1. Population, population elements and stratification

Following guides on sampling, the very first step is to define the population and identify its elements. In the case of reference corpora, the target population is broadly demarcated by geographic borders. In the case of the BNC, the target population is, for example, defined as "the state of contemporary British English in its various social and generic uses" (Aston & Burnard 1998: 28) – so in essence the entirety of language, be it in spoken or written form, produced in Great Britain during a specific time period. Similarly, the target populations of the components of ICE are defined as "the English in each participating country" (Nelson 1996: 28) with a focus on "'educated' or 'standard' English" (Greenbaum 1996: 6). The Brown and the LOB corpus define the population as English prose printed in 1961 in the United States or the United Kingdom respectively (cf. Johansson 1978, Francis & Kučera 1979).

Some authors differentiate three distinct ways of defining the population: "language production, language reception or language as a product" (McEnery et al. 2006: 19–21; Biber 1993: 245). The first two are centred on the individuals who produce or consume language, so that basic demographical details could form the basis for the sampling process (as is to some extent the case for the demographically sampled spoken section of the BNC). In case of the last option, language as a product, McEnery et al. (2006: 19–21) write that the notion of register serves as the basis for defining the population. Corpus linguists seem to agree that we should set our focus on language production rather than reception (see Clear 1992: 24–26 for a discussion).

By recognizing that language varies by situation of use (i.e. register), it follows that we are dealing with a stratified[11] population that is made up of relatively homogenous subpopulations and that those strata are delineated by the notion of register. In a CL context, stratification is discussed under the heading of *balance,* and corpus linguists recognize that the overall representativeness of a reference corpus hinges on the inclusion of a wide range of strata or registers. However, compiling an exhaustive list of all strata and gathering samples from each stratum is not only impractical, but most likely impossible. One route to tackle this problem could be to redefine the population in terms of core and periphery and set our focus on core registers of the population

[11] To avoid terminological confusion: I use the term *stratified* or *stratification* in the context of *population* when I want to highlight the fact that the population is made up of relatively distinct homogenous groups. In contrast, I use *stratified sampling* when referring to a form of probability sampling and *quota sampling* to its non-probability counterpart.

under investigation[12] – although this naturally introduces further methodological issues akin to those when trying to define the core vocabulary of a language.

Since register is a multivariate phenomenon, basing the stratification on one variable alone is not an option. Corpus compilers instead create multi-layered schemes for stratifying the population with different defining variables for each level. These multi-layered and usually hierarchically organized schemes are referred to as *sampling schemes* in CL. Which registers are included in the sampling scheme is often based on the intuition of experts: the sampling scheme of the Brown corpus was decided on by a group of corpus-versed linguists after a conference held at the Brown University in 1963 (Francis & Kučera 1979). Similarly, the design of ICE is the result of discussions among the participants in the project (Nelson 1996: 27). Although the compilation of the BNC is otherwise remarkably transparent, the BNC handbook does not detail how the individual categories for the written and context-governed sections of the BNC were established (cf. Burnard 2000).

Concerning the depth of the sampling scheme hierarchy, mode is the usual top-level distinction if the corpus includes spoken and written material (as is the case for ICE and the BNC). The depth of the sampling scheme as well as the registers included differ from corpus to corpus. As for the Brown and LOB corpora, which include only written material, communicative purpose (i.e. imaginative vs. informative writing) serves as the top-level distinction. The distinctions on the subsequent two levels are not as stringent, as they are partly based on medium (e.g. press), partly on broader subject (e.g. religion) and on literary genre. Similar criticism is also voiced in Leitner (1992), who criticizes the ICE sampling scheme as he sees the hierarchy as well as the individual strata as too fuzzy. In contrast, Lee (2001: 51) writes that it is not problematic if strata are based on different attributes and are of different generality (Lee 2001: 51). In contrast to the definition outlined previously, the registers of these corpora are not defined by a window of acceptable situational characteristics; instead, corpus builders instead relied on a mutual understanding of what the register label refers to.

Although it is practically unavoidable that the definitions of the target populations of reference corpora are fuzzy, let us nevertheless assume for the remainder of this section that we have indeed found a way to define the population unambiguously. While it is then at least conceivable to imagine the target population (i.e. the entire body of language we wish to study), segmenting it into distinct elements already poses the next challenge: Should we split this theoretical body into paragraphs, sentences or even words? Or should more complex units, such as utterances or texts, form the basis for the division? Although the currently favoured terminology in CL speaks of texts,

[12] The admittedly crude idea I have in mind is randomly sampling individuals from a population, monitor their language reception (similar to the methodology described in Wattam (2015)) and, based on the findings, identify the registers an individual is most likely to encounter in daily life. Disregarding all practical issues this would entail for a moment, it would at least provide us with a starting point instead of taking corpus balance as "an act of faith" McEnery et al. (2006: 16). The results then could be used to benchmark or calibrate findings gleaned from currently available corpora.

the notion of a text itself is fuzzy and probably best defined as coherent pieces of discourse. In the case of written pieces, such as newspaper articles, pieces of academic writing or novels, this may be relatively straightforward, but how are we to treat speech events such as an informal conversation among friends or a discussion among politicians broadcasted on TV, for instance?

2.2.2.2. Sample units, sample design and sampling frame

Although Sinclair (2005) suggests including texts in their full length (be it an entire conversation, a novel or just a shorter press editorial), common practice in corpus compilation is to sample a relatively fixed number of words from texts which are longer (usually for practical or copyright reasons) and concatenate multiple texts into one text file if individual texts are shorter. The Brown family and the components of ICE, for example, use 2,000 word text files, while the BNC uses up to 40,000 word text files (depending on the section) (cf. Burnard 2000 on the sample size of the individual sections in the BNC). While Biber (1990) found relative homogeneity in the distribution of text-internal features when segmenting texts into 1,000 word chunks, Nelson (2010: 58–59) identifies potential situations where this still could be problematic and concludes that the sample size should depend on the study goal – in a genre perspective, for example, entire texts are the main focus of the analysis.[13] When adhering to the terminology used in statistical sampling, we sample only parts of the population elements. In such a scenario, the excerpts of texts represent the sample units and should ideally be randomly selected from different positions in the complete texts – as is the case for the written texts in the BNC and the Brown corpus. For ICE, no such information is available. The sampling of excerpts thus in some sense relies on the elements being homogenous in the distribution of the population parameters to be studied (i.e. text-internal features in linguistic terms).

Once the target population is demarcated and it is established what the population elements and sample units are, it is vital to decide upon a sample design and, if probability sampling is intended, define the sampling frame. For compiling reference corpora, corpus linguists widely agree that the sample is to be stratified by register. Although probability sample designs are sometimes treated as the *conditio sine qua non* of representative sampling from a population, they require the construction of a sampling frame and the subsequent random selection of sample units from it. It is easy to see how this becomes problematic in the case of sampling language: The construction of a sampling frame requires a relatively exhaustive list of sample units. While it is unrealistic to expect the sampling frame to cover all potential population elements (resulting in *undercoverage* of the frame population; cf. above), all attempts must be made so that the frame resembles the target population as closely as possible.

The manuals of the BNC and the Brown or LOB corpus give a relatively detailed description how the corpus compilers tried to achieve this and how sampling frames for individual registers were

[13] It must be noted, however, that Biber's analysis of registers has a strong focus on pervasive linguistic features.

constructed. By combining multiple sources such as bestseller lists, catalogues of books published per year and library lending statistics, the compilers of the BNC constructed a sampling frame from which a random sample could be drawn (cf. Burnard 2000: 5–11). In terms of sample design, the books section, for example, represents a mix of random sampling and stratified or quota[14] sampling: around half of the books section was selected randomly from a catalogue that lists books printed in Britain in 1992, the remaining books were chosen manually[15] from library lending statistics, bestseller and prize-winner lists to fulfil fixed quotas of topic areas (*domain*). Burnard (2000: 7) writes that these quotas are "based loosely upon the pattern of book publishing in the UK during the past 20 years or so", but makes no mention of the exact numbers. Elements collected through simple random sampling were post-stratified by the variables *time*, *domain* and *medium*. Elements of the sampling frame which did not fit the inclusion criteria (e.g. non-British writers) were excluded. In a similar fashion, the compilers of the Brown corpus chose a stratified sample design and sampled most strata from the library catalogues of Brown University and the Providence Athenaeum (Francis & Kučera 1979). The manual accompanying the LOB corpus (Johansson 1978) is equally detailed in the description of the compilation process. The author also outlines instances and the reason for when the compilers deviated from simple random sampling (e.g. to favour national press in the newspaper section or to include specific journals in the periodical section). If an exhaustive list of publications is not available for a register, adding additional stages in the sampling frame may be another viable solution. In the case of newspaper writing, this could be realized as randomly selecting newspapers published in a certain region or country at the topmost level, then individual issues and finally individual texts (as was for example done for the Brown corpus, cf. Francis & Kučera 1979).

The design of the spoken section of the BNC is more complex, and the sample design employed can best be described as a form of purposive quota sampling on multiple levels. In contrast to the sample designs outlined above, the sample units are not publications (i.e. books or periodicals), but individuals. The population elements (or the unit of analysis) are still texts (i.e. mostly excerpts of recorded conversations). Crowdy (1993) describes the sampling procedure in detail: On the topmost level, the corpus compilers divided the UK into three supra-regions (North, Midlands and South) from which 12 regions were selected. The spoken part of the BNC is roughly equally divided into two subsections: the *demographically sampled* and the *context-governed* section. For the demographically sampled section, a total of 30 locations were selected based on their ACORN (*A Classification of Residential Neighbourhoods*) profile within the 12 aforementioned regions. Fixed quotas of individuals were then sampled based on their age, social class and gender. The context-governed section was further divided into four domains (*education/informative, business,*

[14] Quota sampling is the non-probability counterpart of stratified sampling and does not require the construction of a sampling frame as it is based on non-probability sampling methods, such as availability sampling. Although Burnard (2000: 10) describes the procedure as systematic, it is unclear whether probability sampling was employed.

[15] Again, no mention is made of what this exactly entails.

public/institutional, leisure), which were in turn further subdivided into monologues and dialogues and these then into several text categories. As most text categories required their own sampling strategies, the description of each is omitted here (see Crowdy 1993 for further details).

In contrast to this, the manuals issued with components of ICE generally only list the sources of the texts included in the corpus, but make no mention of a sampling frame or sample design. Even in the case of ICE-GB, which was designed as a pilot project of ICE (Nelson et al. 2002: 3), the manual does not detail the sampling procedure. For L1 components of ICE, it would be at least conceivable to employ sampling strategies similar to those of the Brown/LOB corpus or the BNC. For varieties where English has the status of a second language, lists of publications might simply be not available. The same is true for the spoken section: A record of individuals speaking English is most likely also not available thus making it next to impossible to employ simple random sampling methods.[16] Another critical problem in the compilation of ICE corpora is that even if lists of publications for a register are available, corpus compilers have to check the background of each individual writer whether they indeed meet the inclusion criteria (i.e. whether it is truly a native speaker of the variety in question). Although it is understandable that due to the scope and nature of the project compilers of ICE components need a certain degree of leeway in the sampling procedure, the lack of transparency is nevertheless disconcerting as it renders assessing the representativeness (and thus also the comparability) of the corpora difficult.

2.2.2.3. Size

As mentioned in the introduction of this chapter, alongside representativeness and balance, size is one of the central concepts in the compilation of a corpus or indeed in the construction of any sample. Yet I have omitted it from the theoretical discussion of sampling as usual ways for determining the required sample size are not applicable in the context of reference corpora. For one, reference corpora by their very nature are not tailored to a specific study goal or to the investigation of a specific population parameter. Instead, reference corpora are meant to be all-purpose databases for linguistic investigations. For another, many corpora are non-probability samples so that a calculation of a required sample size is not theoretically defensible in any case. Whether a corpus is therefore indeed large enough (i.e. whether the sample size is sufficient) has to be determined on a per study basis. This is also underpinned by Biber (1993), who addresses the problem of calculating the required sample size by studying the distribution of some linguistic variables within texts and between texts. According to his findings, the distribution of linguistic variables can vary largely on both levels, which in turn necessitates an adjustment of the sample sizes. Where one might quickly reach the limits when studying less frequent phenomena in smaller corpora such as components of ICE or the Brown family, mega corpora such as the

[16] Surely one could argue that we could employ some form of geographical cluster sampling and simply discard individuals who do not speak English, but this would i) dramatically increase the time and effort necessary for gathering data and ii) it would introduce a whole new set of methodological issues (e.g. non-response bias).

GloWbE (Global Corpus of Web-based English) often come with caveats common to unannotated and automatically collected material (see e.g. Werner 2016: 266–267 on using GloWbE or Schlüter & Vetter 2020 for a discussion on using the Google *n*-grams data). Although there has been a constant push in corpus linguistics towards ever larger corpora, small and carefully compiled corpora are still extensively used and, at least in the case of ICE, new ones created. Mair (2006) divides corpus linguists into two traditions and tries to reconcile those. Followers of the first of these traditions favour said small and carefully compiled corpora (e.g. Brown), whereas the other group focuses on big but messy databases (e.g. web-based corpora). He illustrates how both types of databases can be used complimentarily to investigate the same phenomenon (the *get*-passive) from different perspectives.

2.2.3. Corpus comparability

Comparing data derived from different sources is certainly a highly enlightening endeavour and common practice in nearly all empirical sciences – with linguistics forming no exception. Despite the fact that various sources of data (e.g. experiments, elicitation studies, etc.) could be used for comparisons using data triangulation, the focus of this study lies on the family of comparable ICE corpus components. Where the compilation of reference corpora has already necessitated the discussion and definition of representativeness, the compilation of comparable corpus families complicates this issue further. Sigley (2012: 68) points out that "there is an unavoidable conflict between the requirements of comparability (i.e., that all of our corpora should have parallel content as far as possible) and representativeness (i.e., that each corpus should accurately reflect the variety it is intended to sample)". He argues that, for example, registers[17] might not be used for the same function in different regional varieties or that specific literary genres in a certain variety may not have an equivalent in other varieties (e.g. American Western novels). Concerning comparability, Greenbaum (1996: 5) already noted in his initial description of the ICE project that achieving comparability for such a large number of individual corpora is exceedingly difficult. As distorting factors, he lists different topical foci of the text samples, differences in the biodata of the speakers sampled and the potential unavailability and consequent substitution of certain text categories. Diachronic variation due to time gaps between the releases of individual ICE components (Werner 2014: 112), difference in length of the timeframe of compilation, cultural differences in the interpretation of registers as well as the influence of digitalization can be added to this list (Hundt 2015: 383–386). Hundt (ibid.) further questions the naturalness of the section *private conversations* as, especially for ICE countries where English is used as a second language, data were partially gathered in interview-like settings. Despite this criticism, the ICE family has proven to be an invaluable resource in the study of World Englishes and has been used extensively for comparative studies.

[17] Following the definition of register as proposed in section 2.1.2.2, it is either the conventionalized *register label* that is used differently in different cultures to refer to similar, but slightly different situations of use or the composition of a register in terms of its subregisters that differ between cultures.

Despite these apparent shortcomings, components of ICE are often compared. Theoretically, three types of comparisons are conceivable: i) the comparison of entire corpora (e.g. the comparison of the distribution of, say, the *will* vs. *going to* future in ICE-GB and ICE-CAN), ii) the comparison of similar parts of different corpora (e.g. contrasting the use of discourse markers in private conversations in ICE-GB and ICE-CAN) and iii) the comparison of different sections within one corpus (e.g. the frequency of present perfects in sections of ICE-GB). I will refer to comparisons of type i) and ii) as *between-corpus* comparisons and iii) as *within-corpus* comparisons.

Although I consider within-corpus comparisons as relatively unproblematic as they are basically studies in register variation and components of such corpus families must be used with the same amount of prudence as any reference corpus in any case, a few points still merit discussion. The first issue I would like to address is the variability we encounter in corpora, which is also extensively discussed in Gries (2006). Following up on Schlüter (2006) (who compares corpus-based studies on the frequency of the present perfect, finds high variation between individual studies and consequently questions the reliability of corpus-based findings), Gries (2006) rightly argues that the common practice of reporting frequencies or proportions of linguistic variables without also reporting some measure of dispersion seriously impairs the comparison of results obtained from different (parts of) corpora. In order to identify relevant differences encoded in registers, he proposes to "summarise the frequencies [...] by generating boxplots for every level of granularity suspected to be important" (Gries 2006: 122). The benefits of this approach are evident: For one, the researcher can easily identify if a variable encoded in the registers might have an influence on the parameter of interest, and, perhaps more critically, quantifying the variability helps to put the findings into perspective. If we encounter vastly different means for the parameter in question for different sections of a corpus, but also find high variation within these sections, then the difference in means might not be as significant as it appears. On the other hand, if we find a difference in the means with very little variance in the individual sections, then this might point to a relevant finding. In any case, Gries (2006) exemplifies the importance of taking into account variability within (sections of) corpora (see also Conrad 2015 on register variation).

The second issue I want to discuss is the use of pre-defined text categorizations for *between-corpus* comparisons. While extensively annotated corpora, such as the BNC, allow linguists to go beyond the comparison of these pre-established corpus categories or registers, this kind of information is not included in many corpora or it would entail a substantial amount of cross-referencing the manual accompanying the corpus with the data – as is the case with ICE. Only some components of ICE come with manuals that detail metadata, such as the biodata of the individual speakers or information on the authors of texts. In general, the ICE metadata rarely seems to be included in studies (see Hansen 2018 for an exception). We thus have to rely on the pre-established categorization, order and selection of texts. The sampling scheme of ICE (cf. Table 3 in section 3.1.1) is organized in a hierarchical structure with mode (spoken vs. written) forming the topmost distinction. These sections are then divided into further subsections, ranging, for example, from published material (e.g. academic writing or news writing), unpublished written texts (e.g. letters or student writing) to formal spoken texts (e.g. broadcast discussions) and informal conversations. During the beginning of the ICE project, Leitner (1992) criticizes the established categories in the ICE sampling scheme for being "heterogeneous, ordered unsystematically, and not applied consistently to all category 'paths'" (Leitner 1992: 44). Although he calls for a revision of the ICE sampling scheme to reflect the variable character of text categories, the sampling scheme remains unaltered. I consider this fuzzily defined stratification (see previous section) combined with the lack of annotation of situational characteristics of the texts as highly problematic. It means that the labels of text categories or registers – even though they suggest objectivity and direct comparability – are opaque. One result is that it is neither possible to reorder texts based on specific situational characteristics nor to investigate their influence on a variable of interest. Alternative groupings of texts would have to be based on linguistic, text-internal features (e.g. Sigley 1997, Gries 2006, Sigley 2012). Yet, this might not be possible at times if we are dealing with small differences in register, such as, for example, the circulation and readership of newspapers. The *editorials* as well as the *press news reports* section of ICE-CAN include a wider variety of newspapers (local vs. national, weeklies vs. dailies) than, say, ICE-JA, where the vast majority of texts mainly stems from a single newspaper (cf. Sand 1999: 20). Naturally, this difference in sampled material is a direct reflection of local differences in geography, population, urbanization, the status of English and a number of other factors. The composition of said two corpus sections illustrates the conflict addressed earlier (representativeness vs. comparability). To achieve comparability, the sampling scheme of ICE is based on the same categories of situationally defined language varieties (i.e. registers). Yet, small-scale differences in register are unavoidable if both sections also strive for representativeness. Detecting variation in the situational characteristics of texts, however, would be facilitated if the text-external factors delineating a register were part of the annotation.

Another issue that needs to be addressed in the context of *between-corpus* comparisons is raised by Sigley (2012) and concerns the interpretation of observed differences. Consider the case where two components of ICE are compared, and we find a significant difference with respect to the frequencies of occurrence of a certain linguistic variable. Sigley (2012: 66–67) offers five alternative explanations before such an observed difference should be attributed to *regional* variation:

- systemic differences in stylistic norms,
- enforced institutional norms,
- diachronic drift of registers,
- social distribution of contributors to a register (e.g. more female authors) and
- differences due to the sampling procedure.

In the light of the numerous distorting factors, the question I want to raise here is when do we attribute some observed difference in the distribution of a given linguistic variable to regional variation? Do we only consider global patterns that occur across a wide range of registers and variables as regional variation? How are we to treat the enforcing of different stylistic norms in a certain variety and register? Does this not ultimately lead to variety-specific differences, regardless of whether they are a result of prescriptivism? Similarly, a diachronic drift of genres or registers as observed by Biber & Finegan (1989), who attested a historical drift towards a more oral style for three genres or registers, raises the question of how much a register can change before we consider it a new register. Biber & Finegan (1989: 516), for example, hypothesize that a change in the target audience of the investigated genres or registers due to a spread of literacy might be one of the causes for the observed drift. If we follow this line of argumentation, it could be argued then that the register per se need not have changed, but rather that we are witnessing one register changes into another and that both share the same label. Intended readership is a text-external factor, after all, and registers are defined text-externally.

On a similar note, let us assume that there are indeed compositional differences of a single register between two corpora representing one variety each – say twice as many texts in the register "academic writing" in variety A are authored by multiple researchers compared to variety B. Let us assume further that we investigate self-reference as a linguistic variable, and as possible variants we observe impersonal self-reference (e.g. *one*), first person singular pronoun (*I*) and first personal plural pronoun (*we*). If we now find that variety A uses a higher frequency of *we* for self-referencing compared to variety B, this would be hardly surprising. Although such differences are often readily interpreted as regional differences, without detailed information on the sampling process and a detailed annotation of register variables, such an interpretation is rather daring. Are the samples really representative of the populations, or is sample A biased towards co-authored texts or sample B towards texts with a single author? Even if the samples are largely representative of the individual populations and we have access to the metadata to distinguish between co-authored and single-authored texts, should we interpret the observed differences as regional variation? On the one hand, a language user is likely to encounter more co-authored texts in variety A (and consequently more self-referencing using *we*) than in variety B. On the other hand, we might argue that the observed differences in self-referencing are due to register differences – the number of authors is a text-external criterion. The above considerations spawn a number of uncomfortable methodological questions that unfortunately elude easy answers. For instance, does every change in text-external criteria constitute a new (sub-)register? If the composition of a register in terms of subregisters differs in two varieties and this is also reflected in the distribution of linguistic features, should this be considered regional variation or register variation? In the absence of detailed information of the sampling process, how can we be sure that our samples are representative?

Returning to *within-corpus* comparisons, let us suppose that in the course of comparing the distribution of variants in various registers, certain registers exhibit more variability than others. We might conclude that the linguistic variable is less strictly codified in these registers. An alternative explanation could be that these registers allow for more variation with respect to the situational characteristics that influences the distribution of the variable of interest. Both explanations seem plausible, but not testable with corpus data alone. Registers that exhibit a high variance might be comprised of more distinct subregisters. This ties in with the notion of *granularity*. *Granularity* or *level of granularity* refers to the various hierarchies of sampling schemes of corpora. Studies can be carried out on coarser level of granularities (e.g. only taking into account mode) or finer level of granularities (e.g. taking into account individual registers). Again, the problem I want to point out relates directly to Leitner's (1992) criticism. With merely opaque category or register labels to rely on, it is unclear how broadly or narrowly the individual categories are defined. Surprisingly, Nelson (1996: 29) already identifies this issue during the initial phase of ICE and writes that "studies [...] may reveal that some categories will have to be conflated, or that some will require further sub-classification". While methods for establishing homogenous categorizations of texts have been proposed (see again Gries 2006, especially the approach utilizing bootstrapping), identifying the situational characteristics underlying those groups remains mostly guesswork unless time-consuming qualitative methods are employed.

It is therefore unsurprising that relatively few attempts have been made to empirically measure the comparability of corpora. One exception is Sigley (2012), who compares texts and text categories of the Brown family and the Wellington Corpus of Written New Zealand English (WWC) based on a formality index introduced in Sigley (1997). Apart from this, other approaches to the comparison of corpora are based on word or *n*-gram frequencies (e.g. Rayson & Garside 2000, Kilgarriff 2001, Denoual 2006). It must be questioned, however, whether similarity measures based on text-internal features are indeed helpful in determining corpus comparability. Any quantitative evaluation of comparability depends on how the material was sampled in the first place and whether contextual information is available that allows an interpretation of such measures. Until such measures have been established, it is up to the individual researcher to evaluate on a per case basis whether the data are sufficiently comparable for the purposes of individual investigations.

2.3. Clustering of texts

The classification and clustering of texts has applications well beyond the description of registers and, as might be expected, the literature is vast. The following overview is therefore restricted to studies related to the clustering of text types (i.e. text categories which are relatively homogenous in terms of linguistic features). But even within corpus-based register studies, a multitude of methods has been developed and applied over the past few decades, so that the following overview will concentrate on more established methods.

What all approaches have in common is that the clustering or classification is based on a vector representation (or bag of words/features model) of the texts. That is, each text is sliced into chunks and the frequencies of unique chunks serve as a representation of the text. The studies differ in the methods applied to these representations, the size of the chunks (e.g. the *n* in *n-grams*) and whether some sort of abstraction was applied to the original text (e.g. POS tagging, parsing or other forms of linguistic tagging). As most corpus-linguistic studies focus on discovering structure in the data rather than assigning new texts to predefined categories, researchers commonly employ statistical tools for clustering (mostly hierarchical cluster analysis) and dimension reduction (e.g. multidimensional scaling, principal component analysis or factor analysis). In contrast, studies interested in the discriminatory power of certain feature sets use prelabelled texts as training data for machine learning algorithms and evaluate the accuracy of different feature sets.

Although a myriad of representations of the data are theoretically conceivable, most studies rely either on character *n*-grams (e.g. Cavnar & Trenkle 1994), word *n*-grams (e.g. Gries et al. 2011) or POS *n*-grams (e.g. Tang & Cao 2015) for the clustering of registers (with *n* being usually between 1 and 5). With the goal of describing larger patterns of variation instead of classifying texts, another strand of research follows Biber's (1988) approach and relies on complex linguistic annotation and factor analysis.

Overall, a review of previous studies that employ *n*-grams for the clustering of text types reveals the following global patterns (individual studies referred to here will be discussed in more detail later):

i) More complex representations of the data seem to result in a higher classification accuracy. Complexity here refers to either the length of *n*-grams or, if tagging is used, the complexity of the tagset: Bi-/tri-grams outperform monograms; the CLAWS5 tagset outperforms an impoverished POS tagset (e.g. Santini 2004, Gries et al. 2011, Fang & Cao 2015: 71–82, Tang & Cao 2015).

ii) An increase in the number of subgenres and more fine-grained distinctions between those (also referred to as *level of granularity*) leads to less accurate classification results (e.g. Cavnar & Trenkle 1994, Santini 2004, Petrenz & Webber 2011, Fang & Cao 2015; with the notable exception of Gries et al. 2011).

iii) Text types strongly correlate with registers, i.e. there appears to be a strong link between the distribution of linguistic features and situational characteristics.

Biber's work (esp. Biber 1988, Biber & Egbert 2018, Biber & Conrad 2019) certainly forms a cornerstone in the quantitative description of registers. Using an automated tagging script, he annotated the texts of the LOB and London Lund Corpus (LLC) with 67 linguistic features (cf. Biber 1988). The frequencies of these features then served as the basis for a factor analysis with the goal of identifying what he refers to as *dimensions of variation* (or *latent variables* in statistical terms). The underlying idea is to group co-occurring linguistic features into meaningful categories (factors). His analysis resulted in six factors. Although his focus is not so much on the classification of texts but on identifying global patterns of register variation, his work illustrates vividly that the text-external realm (or situational context) is intrinsically linked to text-internal, linguistic features. What sets his work apart from other methods discussed in this section is the choice of variables – he included features from various linguistic domains instead of merely surface features, such as word frequencies or *n*-grams. Note that while his study provided invaluable insights into register variation, the majority of studies which apply methods for unsupervised clustering rely on much simpler features and still achieve very good results.

Fang & Cao (2015) compare the classification accuracies for three different representations of the texts in ICE-GB: monograms of a fine-grained POS tagset (AUTASYS tagset, Fang 1996), monograms based on an impoverished version of it (similar to the Oxford Simplified Tagset used in the BNC) and word monograms. The three sets of frequency profiles are then split into a training set and a test set, and a naïve Bayes (NB) classifier is used for the classification task at various levels of granularity. The accuracies for the fine-grained tagset are very close to those of the word-based approach, and both outperform the impoverished tagset by far. The exception is when texts are classified by mode (i.e. spoken vs. written): in this case, both POS tagsets outperform the word-based model. The experiments of Fang & Cao (2015) further show a substantial drop in the classification accuracy on finer levels of granularity: they achieve accuracy rates of 99.8% when classifying into spoken/written texts, 74.7% on a genre and a mere 58.2% on a subgenre level (Fang & Cao 2015: 77).

A similar pattern emerges in Santini (2004). She computed frequency profiles based on POS *n*-gram frequencies (for *n* = 1-3) for texts taken from ten registers in the BNC (4 spoken, 6 written). Again, an NB classifier was used. Overall, she reports the highest accuracy rates for the classification task of spoken vs. written texts (98.5%). When classifying texts within these two individual sections, the accuracy for the spoken section is higher (94.6%) than for the written section (78.9%) or when texts of both sections are used (82.6%) (cf. Santini 2004: 4). The relatively high accuracy for the spoken section is likely the result of the choice of registers: There is a split between unplanned and planned discourse for the spoken registers (spontaneous conversations, interviews, public debates, planned speeches) which is not present to this extent in the selection of written registers (academic writing, print advertisements, biography, instructional texts, popular magazines). Tang & Cao (2015) observed similar trends and found that in general, longer *n*-grams outperform shorter ones, and a more complex POS tagset has a higher discriminatory power than an impoverished one. In their experiment, they compared register classification accuracies for the BNC Baby for *n*-grams (*n* = 1-5) based on two tagsets (CLAWS5 vs. Oxford Simplified). Gries et al. (2011) form the exception to the rule. They explore how well word *n*-grams of varying lengths (*n* = 1-5) can represent the original structure of the BNC Baby and found more accurate results at finer levels of granularity.

Sigley (1997) follows another approach: He proposes considering alternative classifications for texts, e.g. based on their formality, as linguistic variables may be more influenced by concepts other than by register or mode (on which the pre-established categories are based). For this, he calculates frequencies of a number of easily measurable features (e.g. frequency of pronouns, hedges, amplifiers, adverbs, core verbs) for texts and subjects these to a principal component analysis (PCA). The first of the two resulting factors he interprets as formality index, the second as related to abstractness.[18] Texts (or registers) can then theoretically be grouped according to this formality index, instead of by the pre-established hierarchy of the sampling scheme.

Naturally, the choice of the method and data representation must be attuned to the given task. Where POS *n*-grams, for example, gloss over differences in topic and quite possibly style, they can capture shallow syntactic structures (especially for $n \geq 3$). Word *n*-grams, in contrast, retain very detailed information, such as collocational preferences and topical focus but could lead to a data sparsity problem for shorter texts or texts with a high type-token ratio. POS *n*-grams should therefore, at least theoretically, be better suited for clustering shorter texts with a high topical diversity than word *n*-grams. If lexical information is important or a topical focus is relevant, word *n*-grams may result in more meaningful clusters. This may, for instance, be relevant when clustering newspaper texts by sections or pieces of academic writing by discipline. Similarly, the decision on whether only a selection or all features enter the analysis is crucial as many traditional statistical tools were not designed for use in a high-dimensional data space (cf. James et al. 2013: esp. section 6.4).

2.4. The changing face of corpus linguistics

While linguists have worked empirically long before the advent of corpus linguistics (cf. Meyer 2009 on the empirical tradition in linguistics or Tognini-Bonelli 2010 for a brief overview), corpora have allowed linguists to support their theories with quantitative data on an unprecedented scale. It is therefore unsurprising that corpus linguistics has received ample attention in the form of lively theoretical and methodological discussions. McEnery & Hardie (2012), for example, critically discuss major trends and schools of thought that developed within the context of CL. Besides contrasting the neo-Firthian and functionalist views on corpora, the authors also attempt to sketch the future path of CL (McEnery & Hardie 2012: 225–227), where a corpus linguist is not a linguist who works with corpora (as seems to have been the case since the beginnings of CL), but a linguist who works on the theory behind the compilation and use of corpora and develops new corpus-linguistic methods for other linguists to use. In the future projected by McEnery & Hardie (2012), the use of corpora will lose its unique status and CL will be simply yet another tool in the empirical linguist's toolbox.[19] This change can already be witnessed in the way corpora are almost casually treated as one among many sources of evidence for linguistic insights and are often used alongside experimental data or data from elicitation experiments – although there appear to be differences between the various linguistic disciplines: Gilquin & Gries (2009) found that while psycholinguists seem to readily compare experimental results to corpora, corpus linguists (in the traditional definition) appear to rather reluctantly incorporate other kinds of data in their studies.

[18] However, he cautions that many of the elements scoring on the second factor are multifunctional and any interpretation therefore must remain tentative.

[19] It is strange enough that we talk about a corpus linguist, but not about questionnaire linguists or an interview linguist. This already hints at the special role corpus linguistics seems to play and that corpora have a status that goes beyond a mere source of data.

Another result of this changing face of the work of a corpus linguist is the continuous development of software such as the freely available concordance programs *The Corpus Workbench* (Evert & Hardie 2011) and its *BNCweb*-inspired web interface *cqpWeb* (Hardie 2012), *AntConc* (Anthony 2018) and *shinyConc* (Wolk & Fastrich 2019). While this list is by no means exhaustive, it serves to illustrate how (corpus) linguists develop and release tools for their (maybe also less technically adept) colleagues. *The Corpus Workbench* (CWB) is certainly one of the most powerful tools for managing and querying corpora available, but requires a substantial amount of technical expertise. It consists of tools for indexing and encoding corpora and their metadata into a special file format that allows for the efficient querying of the corpora. In conjunction with a webserver and *cqpWeb*, it empowers many corpus websites of linguistics departments (e.g. https://cqpweb.lancs.ac.uk/). *AntConc*, on the other hand, seeks to serve another audience entirely. As a lightweight program, which can be run locally on all major operating systems, besides the proprietary *Wordsmith Tools*, it is the perfect tool for handling small corpora. It includes a range of functions, including a concordancer, *n*-grams, collocation analysis and (key-)word lists. Since *AntConc* is not equipped to handle metadata, the recently released *shinyConc* seeks to close this gap. *shinyConc* is a concordancer written in the programming language *R* (R Core Team 2019) and the web framework *shiny* (Chang et al. 2019) and allows the user to load self-compiled corpora and filter by metadata.

It is in this spirit that this study interprets the work of a corpus linguist. This study not only seeks to explore methods for assessing the comparability of components of ICE, but also to make these methods available in form of a computer program (*ICEtree*), which requires no background in programming. This view is admittedly somewhat at odds with Gries (2011: 92–94), who endorses the idea of learning (statistical) programming languages as part of linguistic training. His line of reasoning is that linguists should not be limited by the tools available to them. Without resorting to programming languages, one is limited, for example, to measures of collocational strength that are already part of the software one is using. Similarly, if the corpus software does not support regular expressions, extracting more complex patterns from a corpus might not be possible. These limitations effectively limit the range of research questions a linguist can investigate. Although I generally agree with Gries's assessment that programming languages widen the scope of what we can do with data, learning a programming language requires extensive training and practice. Even after one has mastered a programming language, the writing of scripts or programs is a mentally taxing task. Especially in explorative stages of research, where a researcher may be interested in quickly changing between multiple perspectives on the data, pre-made interfaces and programs offer the advantage of not distracting from the data. Schlüter & Vetter (2020) tackle this problem by using programming languages to code an interactive data visualization that is tailored to a specific research question. In a similar vein, Vetter (to appear) uses an earlier version of the program *ICEtree* for investigating differences in the composition of some components of ICE.

2.5. Interim summary and critical reflection

In the course of this chapter, I defined relevant terms and concepts, outlined the fundamentals of sampling theory relevant to the compilation of reference corpora and tried to apply the terminology to the compilation of some commonly used corpora. In addition, I critically addressed the issue of corpus comparability and reviewed approaches to the clustering and automatic classification of texts.

The terms *text category* and *text variety* serve as theory-neutral terms for referring to groups of texts. If the grouping of texts is based on text-internal criteria, these are referred to as *text types*. In contrast, both *genre* and *register* are defined as text-externally demarcated groupings of texts. The key difference is which theory of categorization is employed. Register utilizes the classical way of categorization, with clear-cut boundaries and an equal status of all members of a category, whereas genres exhibit a prototypical structure with fuzzy boundaries and categories structured around exemplars.

Concerning the sampling strategies of existing corpora, the written section of the BNC, the Brown as well as the LOB corpus mostly rely on a multi-level stratified random sample design, sometimes with manual selection of prominent sources (e.g. in the newspaper section). Where random sampling was not possible, purposive sampling was employed (e.g. for *Skills and Hobbies* and *Popular Lore* in Brown). The randomly sampled parts of the written section of the BNC were post-stratified. The spoken section of the BNC employs a multi-level region-based purposive quota sample design. Due to the lack of descriptions of the sampling processes, it must be assumed that the spoken as well as the written sections in ICE rely entirely on non-probability sampling techniques (purposive quota sampling).

Although basing the stratification on the notion of register is the *de facto* standard for reference corpora, the current practice has a few shortcomings. I consider the definition of register (or better the lack thereof) as one of the most critical since this has not only an impact on the sampling of material, but also on the interpretation of the results gleaned from corpora. And although Biber (1993) discussed this issue and more in his seminal article on representativeness in corpus design, many of his suggestions for improvement have not found traction in modern CL. He argues for stratified sampling and clearly defining the sampling frame (and by proxy the population) in order to ensure representativeness, but writes that ‘a fully representative corpus cannot be determined at the outset’ (Biber 1993: 256) as the existing registers of a speech community need to be determined and the distribution of linguistic features investigated. He consequently argues for a cyclical corpus design where the results of linguistic investigations result in the appropriate modifications of the corpus design. For defining the strata of the sampling frame, he further advocates sampling according to a simplified list of situational characteristics (Biber 1993: 245). While the latter would certainly allow linguists to factor in situational characteristics into multivariate analyses instead of having to rely on opaque register labels, this practice has not caught on (see also the discussion on corpus comparability in this chapter). A bottom-up approach to this problem, i.e. a charting of all registers in a given language variety or speech community, similar to what Biber et al. (2015) exemplify for internet registers, is (still) practically not feasible. For written material, it would be at least theoretically conceivable to gather a list of all material published in a given period of time, draw a random sample, categorize it by situational characteristics and design the sampling scheme accordingly (similar to how the sampling scheme for the Brown and LOB corpora were defined). However, the same cannot be done for spoken and unpublished language. We cannot estimate, for instance, how many conversations are carried out between two, three or more participants, or how many are single-sex or mixed-sex conversations. How much language is produced in a formal and how much in an informal setting? Although some progress has been made using life-logging (Wattam 2015), this issue is still not resolvable in the foreseeable future — not least due to practical constraints — the currently favoured course of action in corpus compilation is designing the sampling scheme in a top-down fashion by experienced linguists. While this appears to be the most practical and plausible approach, we must acknowledge that "any claim of corpus balance [and thus also representativeness] is largely an act of faith rather than a statement of fact as, at present, there is no reliable scientific measure of corpus balance" (McEnery et al. 2006: 16).

Another resulting issue of the current practice is that some corpora (especially ICE) or parts of such, however balanced they may appear, are based on non-probability sampling, and many texts by design have a zero probability of being sampled. Consequently, we must be aware of the limitations that apply to the generalizations we draw from those corpora. If we study a corpus that contains only a few written registers, it seems only logical that our findings apply to the selection of written registers of the language variety under investigation. Findings gleaned from components of ICE, however, are often treated as being representative of the entire language variety.[20]

While the above considerations were voiced before (e.g. Meyer 2004a: 348–350), this practice seems to be regarded as a fair trade-off between theoretical precision and practicality. The same naturally applies if we employ statistical methods that operate under the assumption that the sample was gathered with a form of probability sampling (see Köhler 2013 for a critical and rather bleak evaluation of applying statistical methods in CL). In light of limited resources and practicality, Woods et al. (1986: 55) suggest treating any sample as if it were based on a probability design and judge on a per study basis whether the actual sample design may have an effect on the results. Similarly, Leitner (1992: 43) doubts whether probability sampling is indeed necessary and leads to better results.

An additional problem already outlined in this chapter that further complicates between-corpus comparisons is the lack of transparency in terms of sampling procedure and the definition of

[20] Of course, it seems to be understood in the linguistic community that the databases and resulting findings have to be taken with a grain of salt.

registers. Where some corpora (most notably the BNC and the Brown and Frown corpora) come with extensive descriptions of how the data were gathered, the sampling procedure of many components of ICE remains rather obscure. Concerning the register-based stratification, components of ICE might share the same register labels, but these may still contain different (sub-)registers (see case studies in chapter 5 and the discussion on stratification and corpus comparability). Studies employing bottom-up clustering methods that support prevalent register classifications notwithstanding (e.g. Gries et al. 2011, Fang & Cao 2015), my own studies highlight that it would be naive to rely on these opaque register labels. This is by no means meant as criticism of the respective ICE components. By its very nature, an ambitious project such as ICE has to accommodate for many theoretical and practical constraints (again, see previous section on sampling) and such discrepancies between components of ICE are practically unavoidable. The lack of transparency, however, is avoidable.

2.6. Scope and aim of the present study

The overarching goal of this study is threefold: i) to outline shortcomings of the current practice of ICE with regard to comparability of individual components; ii) to explore quantitative methods that help assess the comparability of some already published components of ICE; and iii) to release a user-friendly program (*ICEtree*) that invites other researchers to scrutinize the presented case-studies and apply the methods described in this study to other parts of ICE not investigated here.

With respect to i) and ii), the individual sections of this chapter reviewed the compilation of reference corpora from a sampling-theoretical perspective and detailed the additional layer of complexity the compilation of comparable corpora adds. Although some of the issues are not resolvable, or at least not within reasonable ethical and practical boundaries, the lack of transparency is disconcerting. Further, earlier criticism has remained largely of a theoretical nature and not been investigated empirically. To this end, four case studies (chapter 5) are presented that exploit the strong link between situational characteristics (registers) and concrete realizations of texts (text types) to point out where the lack of transparency in corpus compilation leads to differences in register that, in turn, lead to a decrease in comparability. The case studies rely on quantitative, bag-of-words representation of the individual text files of ICE and utilize a range of statistical tools for investigating patterns and differences between (groups) of texts. A detailed description of the data as well as the methodology will be given in the next chapter. The first case study scrutinizes the *press editorials* section of ICE-CAN, ICE-GB and ICE-JA, and finds that this register contains various combinations of subregisters in the individual components of ICE. In order to rule out the possibility that the observed differences are due to sampling errors or chance, the individual components are complemented with additional material. The scope is then broadened by comparing these results to other components of ICE (ICE-USA, ICE-HK and ICE-IND). The following three case studies investigate the sections *skills and hobbies, business transactions* and *phone calls* of ICE-CAN, ICE-GB, ICE-JA and ICE-IND and exemplify how a difference in situational characteristics can be detected with quantitative methods and how these differences manifest in the texts. Additionally, all case studies demonstrate how the analysed texts can be annotated with register variables. The four case studies utilize the program *ICEtree,* which I have developed for this thesis. Chapter 4 describes the program and its functions in detail.

3. Data and statistical methods

This section details the data (3.1.) and outlines the statistical methods (3.2) employed in the case studies in chapter 5 and available in *ICEtree*. Section 3.1.1 provides an overview of the corpora – including a brief history of ICE, which components are used and how they are structured (3.1.1.1). As initial explorations of the corpora with *ICEtree* suggested differences in the composition of some registers between some components of ICE, I reviewed the texts of these sections and grouped them based on defining situational characteristics. Due to the emerged compositional differences, expanding some components with the subregisters not present in the respective component became necessary to increase comparability. Details are presented in section 3.1.1.2.

As hinted at in the introduction, the quantitative approaches employed in this study are based on frequency profiles of texts. The compilation of these frequency profiles or vector representations required extensive data pre-processing, including the manual correction and subsequent removal of mark-up included in the original ICE files and the annotation with a second set of linguistically motivated tags (following Biber 1988). These steps are described in section 3.1.2.

3.1. Data

3.1.1. The International Corpus of English

The project International Corpus of English was launched soon after Sidney Greenbaum presented a proposal to create an "international computerized corpus of English" to complement the then and now still widely used and successful Brown corpus family (cf. Greenbaum 1988). The proposal was met with enthusiasm by the linguistic community, and numerous research teams worldwide announced their support for this project (cf. Greenbaum 1991). At the time of writing, almost thirty years after the project was initiated, ICE has turned out to be a major success: With no less than 15 corpora finished and another 10 in compilation (cf. Kirk & Nelson 2018: 712–714), the project provides the community with invaluable resources for comparative studies of World Englishes. Despite the criticism voiced in previous sections, results obtained from these corpora have without a shadow of a doubt greatly furthered our understanding and knowledge of English varieties and experience in compiling corpora. This is also reflected in the recently published review of the ICE project (Kirk & Nelson 2018), which is based on the results obtained from a written questionnaire that was issued to ICE teams in 2016/2017. The goal of the questionnaire was to elicit the challenges, problems and future developments of the ICE project as perceived by members of ICE teams. Among other things, the review concludes that future components of ICE (also referred to as second generation corpora) should mirror the compilation practice of the older components as closely as possible to ensure comparability. Comparability in general is a central theme in the review and addressed in numerous contexts, most of which are connected to the criticism voiced in section 2.2.3 (i.e. time gaps between components & diachronic variation, differences in sampled registers & cultural interpretation of registers, annotation and tagging). Members of the ICE community who participated in the questionnaire agree that ensuring comparability is of paramount importance and that special emphasis should be placed on the homogenization and regularization of (future) components with regards to compilation and annotation practices.

3.1.1.1. Sampling scheme and included components

To increase the range of applications, the program *ICEtree* includes seven components of ICE, not all of which are analysed in the case studies presented in chapter 5. The included varieties are: ICE-CAN, ICE-GB, ICE-HK, ICE-JA, ICE-SIN, ICE-USA and ICE-IND. The components were mostly chosen for practical reasons as all of these corpora follow the original ICE guidelines more closely than some of the other components (see also Kirk & Nelson 2018: 698 on that matter) and are freely available in a POS tagged version (CLAWS7 tag set, see also the respective section in the appendix). Most of the material in the corpora dates from the early 1990s, even for components that include material from as late as 2008 (e.g. ICE-USA or ICE-JA, see Table 2). Except for ICE-USA, which contains only the written sections, all other components are completed in terms of included sections. It is common practice that texts (or better: text files) in ICE are usually a concatenation of multiple individual texts and comprise approximately 2,000 words each. With 500 text files, each component totals one million words. All national components follow the same hierarchical sampling scheme (Table 3 below) so that the texts are first grouped into spoken/written material, then into public or private material in case of spoken texts and into printed or non-printed material for written texts, then into 12 registers and finally, on the lowest level of granularity, into 32 subregisters.

Component	Material sampled from	Source
ICE-GB	1990 - 1993	Nelson et al. (2002: 5)
ICE-CAN	1988 - 1999	Corpus Headers
ICE-JA	1990 - 2008	Corpus Headers
ICE-HK	1990 - 2004	Corpus Headers
ICE-IND	1989 - 1999	Corpus Headers
ICE-SIN	1989 - 1996[21]	Ooi (1997)
ICE-USA	1990 - 2008	Corpus Headers

Table 2: Date range from which material was sampled for the included ICE components.

[21] No corpus headers are available. Ooi (1997) is based on a conference paper that was presented at ICAME 17 in 1996 and writes that the study is based on the finished corpus, so the material must have been collected between 1989 and 1996.

SPOKEN (300)	Dialogues (180)	Private (100)	Face-to-face conversations (90)
			Phone calls (10)
		Public (80)	Classroom Lessons (20)
			Broadcast Discussions (20)
			Broadcast Interviews (10)
			Parliamentary Debates (10)
			Legal cross-examinations (10)
			Business Transactions (10)
	Monologues (120)	Unscripted (70)	Spontaneous commentaries (20)
			Unscripted Speeches (30)
			Demonstrations (10)
			Legal Presentations (10)
		Scripted (50)	Broadcast News (20)
			Broadcast Talks (20)
			Non-broadcast Talks (10)
WRITTEN (200)	Non-printed (50)	Student Writing (20)	Student Essays (10)
			Exam Scripts (10)
		Letters (30)	Social Letters (15)
			Business Letters (15)
	Printed (150)	Academic writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Popular writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Reportage (20)	Press news reports (20)
		Instructional writing (20)	Administrative Writing (10)
			Skills/hobbies (10)
		Persuasive writing (10)	Press editorials (10)
		Creative writing (20)	Novels & short stories (20)

Table 3: The sampling scheme of components of the International Corpus of English. Numbers in brackets refer to the number of text files in the respective category.

3.1.1.2. Reclassification of texts and corpus expansion

The motivation for the reclassification of texts is the result of initial explorations of the similarities of the POS profiles. Although for most registers, texts formed relatively homogenous clusters, it was for *press editorials, skills and hobbies, business transactions* and *phone calls* where distinct patterns emerged. Close reading of the texts suggested that register differences are the cause for the observed patterns. For this reason, I manually reclassified texts in these sections in some ICE components into distinct subregisters. After the reclassification, I deemed it necessary to further expand the *press editorials* section to achieve a more balanced representation of the newly established registers. For the case study on press editorials, texts in all included ICE varieties were manually reclassified into personal and institutional editorials. A detailed description of the individual types of editorials is presented in section 5.2.1.1. Texts in ICE-GB, ICE-CAN, ICE-JA and ICE-IND the *skills and hobbies* section were manually reclassified according to communicative purpose into instructional texts, informative texts and cooking recipes. *Business transactions* in ICE-GB, ICE-IND, ICE-JA and ICE-CAN were divided into three distinct subregisters based on the interactivity of the texts and whether speakers represent individuals or institutions. The category *phone calls* in the same four varieties was split into private and public phone calls.

Press Editorials

The reclassification of the press editorials into institutional and personal editorials was partly performed by reviewing facsimiles of the texts and partly by checking against current issues of the newspapers for stable patterns. The *press editorials* section of each national component comprises 10 texts of approximately 2,000 words each. The metadata included in the corpora indicate that each text is again a concatenation of one to five texts (referred to as subtexts). This is frequent practice in ICE as texts in some text categories are often shorter than 2,000 words, the standard target for each text unit. When text files include more than one text, the subtexts were mostly taken from the same source and timeframe and in case of *press editorials* also mostly from the same subregister of opinion pieces. For ICE-CAN, ICE-JA and ICE-GB, I reviewed facsimiles of the texts for the manual classification into institutional and personal editorials. For the remaining ICE components (ICE-USA, ICE-IND, ICE-HK), I referred to the metadata and reviewed current issued of the sampled newspapers. The texts in ICE-SIN could not be reclassified as there are no metadata available, thus making it impossible to determine the subregister text-externally. Although both types of opinion pieces share many situational characteristics, they differ significantly in some respects with addressor being most discriminatory. The major differences in the situational characteristics are summarized in Table 4.

Subregister Situational characteristic		**Institutional editorials**	**Personal editorials**
Participants:	Addressor	Institutional / unidentified	Single
	Addressor; social characteristics	Professional journalists, often editors, highly specialized	Often non-journalists or specialized column writers, for *L2Es*: readers
Relationship among participants:	Personal relationship	Distant/strangers	Depends on type of *PE*; for columns closer relationship between reader and writer
	Shared knowledge	Specialized	Specialized; for columns also often personal
Communicative purpose		Report, persuade, argue	Range of purposes: report, persuade, argue, entertain, reveal self
Topic		Mostly political	Recent event, daily topic, for columns also personal topics, for *L2E*: article from previous issue

Table 4: Differences in situational characteristics of institutional and personal opinion pieces.

The aforementioned facsimiles were either provided by the corpus compilers[22] or were retrieved from online newspaper archives (e.g. newspapers.com, newspaperarchive.com or the Google Newspaper Archive). Where facsimiles were not available, the register distinction was inferred based on close reading and research on the newspaper and the author of the article.

Table 5 provides an overview of the manual subregister classifications. While in some cases genre markers (cf. Biber & Conrad 2019: 55) provided clues to categorizing texts as institutional or personal editorials (e.g. reference to earlier newspaper issues in letters to the editor), the main criteria I relied on answered the following questions:

- Is the text labelled as belonging to a specific subregister (e.g. institutional editorials in *The Jamaican Gleaner* or *The Observer*)?
- Does the text include information on the author (e.g. a name, a picture or, in case of external authors, their regular occupation)?
- Is/was the author employed at the newspaper and published regularly?

All texts in ICE-GB were categorized as institutional editorials as these texts do not include information on the author, are accompanied by a logo or a banner of the newspaper or are labelled as editorials. In the case of ICE-JA, all texts were categorized as personal editorials. The metadata of the Jamaican texts include the name of the author, and institutional editorials in the sampled

[22] I am greatly indebted to Bas Aarts, Sean Wallis and John Newman.

newspapers do not feature an author. Additionally, some of the authors are listed as columnists on the website of the newspaper (http://digjamaica.com/columnists, last accessed on 26/09/2018). While not part of the corpus, institutional editorials published in *The Gleaner* are titled 'The Gleaner Editorial' and include no information on the author. The institutional editorials further contain the disclaimer 'Opinions on this Page [sic], except for those in the Editorial above, do not necessarily reflect the views of the GLEANER'. Although only facsimiles of the texts W2E-003, W2E-004, W2E-008, W2E-009 and for parts of W2E-001 and W2E-007 could be retrieved, this pattern was found to be stable when reviewing other issues of this newspaper. ICE-CAN, on the other hand, includes a range of different opinion pieces. Facsimiles for all texts were provided by the ICE-Canada team. Only texts W2E-002 and W2E-008 were classified as institutional editorials. The remaining texts were classified as personal editorials as they are either comprised of columns (W2E-001, W2E-003), comments (W2E-005, W2E-006, W2E-007, W2E-010), or letters to the editor (W2E-004, W2E-009). In the cases where the text files contained a mix of opinion pieces (e.g. W2E-004 or W2E-006), the file was classified based on which register constitutes most of the text file. I classified all texts in ICE-HK and ICE-IND, and most in ICE-USA as institutional editorials on the grounds that the otherwise meticulous metadata do not include information on the author. A review of current and older issues of the newspapers revealed that the global pattern of authorship indication in the subregisters also holds for these varieties. The texts ICE-USA:W2E-003 and ICE-USA:W2E-004 were classified as personal editorials as the newspaper these texts were sampled from do not feature institutional editorials, and some texts include a mention of the author.

As most components primarily contain either institutional editorials or personal editorials, it was necessary to extend the existing corpus sections by the respective missing subregister to allow for a meaningful comparison. Hence, additional texts were collected, formatted according to the ICE guidelines, tagged and converted to profiles containing the frequencies of the tags (see below). This step was only performed for the varieties where facsimiles could be retrieved (i.e. ICE-GB, ICE-CAN & ICE-JA). The texts were selected from the newspapers included in the original corpora and sampled from the same time period to minimize distorting factors. For ICE-GB, an additional five texts containing personal editorials were sampled and labelled GB:W2E-011 - GB:W2E-015. For the Canadian and Jamaican component, the same was done for institutional editorials.[23]

[23] The metadata of the newly sampled texts and subtexts are included in *ICEtree*, the appendix and are also available in the OSF repository of *ICEtree*.

Files	Subregister
CAN_W2E-001, 003-007, 009, 010	Personal_Editorial
CAN_W2E-002, 008	Institutional_Editorial
GB_W2E-001-010	Institutional_Editorial
JA_W2E-001-010	Personal_Editorial
HK_W2E-001-010	Institutional_Editorial
IND_W2E-001-010	Institutional_Editorial
USA_W2E-001, 002, 005-010	Institutional_Editorial
USA_W2E-003, 004	Personal_Editorial

Table 5: Reclassification of texts in *press editorials* in ICE-GB, ICE-JA, ICE-CAN, ICE-IND, ICE-HK & ICE-USA.

Skills and Hobbies

As with the *press editorials* section, *skills and hobbies* comprises 10 texts of 2,000 words each, and individual texts can be a concatenation of subtexts. The composition of the *skills and hobbies* section is potentially much more diverse due to its more fuzzy, exemplar-based definition. Nelson (1996: 33) delineates this section from the second register (*administrative writing*) of the section *instructional writing* and writes that "[t]exts in the skills and hobbies category also offer instruction, but [...] are directed towards a smaller and more specialized readership [...], include publications such as car manuals, cookery books, and gardening manuals [...] [and] often present instructions in a step-by-step format [...]". Although the definition mentions the readership and communicative purpose of the texts as defining situational characteristics, it is only the enumeration of exemplars that provide the reader with a clearer picture of the intended register. It is therefore probably unsurprising that different ICE teams interpret the boundaries of this register differently and that a wide range of sources was used to fill this section. The header files of ICE-IND indicate that the text files ICE-IND:W2E-011-014 contain articles from *The Hindustan Times*, one of the largest English newspapers in India. The remaining texts in ICE-IND are sampled from periodicals such as *Femina* (ICE-IND:W2E-015-016), an Indian women's magazine, *Intensive Agriculture* (ICE-IND:W2E-017-018), a government-issued periodical, and the periodical *Amateur Photography* (ICE-IND:W2E-019). The source of text ICE-IND:W2E-020 is not indicated in the header files. In contrast, texts in ICE-CAN were sampled exclusively from periodicals with narrow topical foci and texts in ICE-GB are excerpts from instructional monographs. The available metadata of ICE-JA were not sufficient to identify the exact source and type of publication of the included texts.

Close reading of texts in *skills and hobbies* in ICE-IND, ICE-JA, ICE-CAN and ICE-GB allowed me to group the texts by communicative purpose into informative and instructional texts (see Table 6). The major difference between the two groups is that informative texts make no attempt to outline how a certain task is best done. Texts in the instructional category, on the other hand, focus on a very specific task (e.g. tips on gardening or animal breeding) and either directly or indirectly instruct the reader on how it is best performed and where potential pitfalls are. Although instructional in nature, I treated recipes separately from other instructional texts as they often consist only of brief imperative statements and exhibit a high frequency of numbers for specifying quantities (e.g. "add 2 cups of sugar"). In the rare instance where one text file included more than one subregister, the text file was classified based on which subregister constitutes most of the text file.

Files	Subregister
CAN_W2D-011, 015-018, 020	skillsHOBBIES_informative
CAN_W2D-012-014, 019	skillsHOBBIES_instructional
GB_W2D-011-020	skillsHOBBIES_instructional
IND_W2D-011-014	skillsHOBBIES_informative
IND_W2D-015, 016	skillsHOBBIES_recipe
IND_W2D-017-020	skillsHOBBIES_instructional
JA_W2D-011, 013-014, 016-020	skillsHOBBIES_instructional
JA_W2D-012, 015	skillsHOBBIES_recipe

Table 6: Reclassification of texts in *skills and hobbies* in ICE-GB, ICE-JA, ICE-CAN & ICE-IND by communicative purpose.

Business transactions

This section again consists of 10 texts of 2,000 words each. Similar to the ICE category *skills and hobbies*, *business transactions*, a subcategory of *public dialogues*, is not clearly delineated. Nelson (1996: 31), who also refers to this category as *business meetings*, writes that the "topic of discussion is known in advance by all the participants" and that "there may even be a written agenda or schedule to guide the proceedings". The handbook of ICE-GB, the first released ICE component, only provides examples of possible subregisters: "business meetings, faculty meetings, and consultations with professionals" (Nelson et al. 2002: 6). It is therefore not surprising that the composition of this category in terms of subregisters differs between components. Cross-referencing with the metadata shows that ICE-GB contains a relatively wide range of subregisters, including professional consultations (e.g. between an architect and their clients), faculty meetings, business discussions and committee meetings. For texts in ICE-CAN, only metadata on the speakers are provided, and it is indicated that only two speakers engage in the dialogues. For ICE-IND, the situation is different: the corpus headers included indicate that the texts contain only meetings of boards and committees of educational institutions. The metadata further detail that some texts contain the reading of minutes of previous meetings or are likely to contain otherwise longer stretches of monologues (ICE-IND:S1B-076 is a Ph.D. viva voce). Through close reading of all texts, I found that there is a steep cline of interactivity present in the texts. Although all texts are labelled as meetings, some are highly interactive with participants engaging in lively discussions and addressing each other, whereas others contain long, monologic stretches where participants present information (e.g. the reading out of the minutes of a previous meeting or the presentation of the agenda), and the remaining participants only respond with very brief comments or interrupt for clarifications. With respect to ICE-JA, metadata indicate that the texts ICE-JA:S1B-071 & ICE-JA:S1B-072 are sales conversations, whereas the remaining texts are wage negotiations between the University of the West Indies (UWI) and the West Indies Group of University Teachers Union (WIGUT). The metadata further contain a note that these texts are verbatim notes, which suggests possible differences in transcription practices. And indeed, close reading revealed that no hesitation markers and interjections were transcribed for these texts. It further stands out that the speakers do not represent individual opinions, but argue in the name of either the WIGUT or the UWI, and that these discussions are highly planned.

Following these observations, I reclassified the texts in *business transactions* into three distinct groups (see Table 8): the first is labelled *businessTRANS report* and contains most Indian texts that show a very low interactivity. The second is labelled *businessTRANS discussion*, contains the British, Canadian, two Jamaican and two Indian texts and is characterized by high interactivity (i.e. high rate of turn-taking and interruptions). The third group is labelled *businessTRANS institutional* and is comprised of the Jamaican wage negotiations. These texts are similar to the previous category; however, the speakers do not represent themselves but rather speak on behalf of institutions. The subregister *businessTRANS report* and *businessTRANS institutional* are characterized by a higher degree of planning compared to the more interactive category *businessTRANS discussions*. The arguments presented in wage negotiations, although uttered in real time, are prepared in advance. Similarly, the minutes of previous meetings to be read out consist at least of notes. The three categories differ in their communicative purposes. Although it is difficult to assign a single purpose to these texts, the main purpose of texts in businessTRANS report is to report, whereas the main purpose of the utterances in the other two sections is either to inform attendees about certain circumstances or to persuade the addressees of a point of view. In terms of situational characteristics, these three groups thus differ in terms of interactivity, the addressors, the production circumstances (i.e. planning), as well as the general communicative purpose. The differences are summarized in Table 7.

Subregister Situational Characteristics		**businessTRANS report**	**businessTRANS discussion**	**businessTRANS institutional**
Participants	Addressor	Single	Single	Institutional
	Addressee	Single, group of individuals	Single, group of individuals	Group, institutional
Interactivity		Low	High	Medium - high
Production circumstances		Planned / real time	Real time	Planned / real time
General communicative purpose		Report	Range of purposes, mainly to inform and persuade	Range of purposes, mainly to inform and persuade

Table 7: Differences in situational characteristics for subregister in *business transactions*.

Files	**Subregister**
CAN_S1B-071-080	businessTRANS_discussion
GB_S1B-071-080	businessTRANS_discussion
JA_S1B-071, 072	businessTRANS_discussion
JA_S1B-073-080	businessTRANS_institutional
IND_S1B-071, 075	businessTRANS_discussion
IND_S1B-072-074, 076-080	businessTRANS_report

Table 8: Manual reclassification of texts in *business transactions*.

Phone calls

Phone calls, a subcategory of *private dialogues,* is also comprised of 10 texts à 2,000 words. Although these conversations "may be overheard, speakers only address each other [...] [and] do not speak for the benefit of anyone else who may be present" (Nelson 1996: 31). The metadata of ICE-GB include information on the relationship between speakers. With the exception of two subtexts, the speakers are either friends or relatives. For ICE-CAN, no such information is available, but the texts themselves are informal and suggest a fair amount of shared personal knowledge, so that it is assumed that these are also private conversations among friends. Based on close reading of the conversations and the topics discussed, most phone calls in ICE-IND appear take place in a professional setting, and speakers are familiar with each other. Except for ICE-IND:S1A-099 and ICE-IND:S1A-100, participants are described as equals in the metadata. According to the metadata of ICE-JA, on the other hand, the Jamaican phone calls are calls to hotlines and radio shows and as such more akin to public conversations. The greetings at the beginning of these calls, where the name of the show (e.g. "Public Eye") or the hotline is mentioned, additionally serve as register markers.

Subregister Situational Characteristics		**phoneCALLS private**	**phoneCALLS public**
Participants	Audience	No	Yes
Relationship among participants	Social role	Equal	Difference in power
	Relationship	Friends, relatives, colleagues	Strangers
	Shared knowledge	Personal	?

Table 9: Differences in situational characteristics between private and public phone calls.

Based on this information, I reclassified texts in *phone calls* into public phone calls *(phoneCALLS_public)* and private phone calls *(phoneCALLS_private)*. Table 10 details the reclassification. These two subregisters differ in a few situational characteristics related to the participants of the conversation (see Table 9). One of the most striking differences is certainly that public phone calls are broadcast so that the speakers are aware that an unenumerated audience is present. The relationship between the participants also differs for both subregisters. It could be argued that the host of the hotline or radio show is in a position of power, as the guests often call to seek counsel. Further, the participants are likely strangers and thus share no personal knowledge.

Files	**Subregister**
CAN_S1A-091-100	phoneCALLS private
GB_S1B-091-100	phoneCALLS private
JA_S1B-091-100	phoneCALLS public
IND_S1B-091-100	phoneCALLS private

Table 10: Manual reclassification of texts in *phone calls.*

3.1.2. Data preparation

All quantitative methods employed in the present study and *ICEtree* are based on frequency profiles. Two types of frequency profiles are available: the first is based on POS tags, the other on linguistically motivated tags (following Biber 1988). The creation of these frequency profiles requires extensive preparation of the data. The preparation process is depicted in Figure 2 and can be broadly separated into two subsequent steps: i) data cleansing, i.e. the manual correction and automatic removal of mark-up, and ii) the conversion of tagged texts into a vectorized representation or frequency profile. For both profile types, the parts-of-speech (POS) tagged versions of ICE-CAN, ICE-JA, ICE-USA, ICE-SIN, ICE-IND & ICE-HK formed the basis. For consistency with the other components, ICE-GB was converted from its syntactically parsed format into a plain text version[24] and then POS tagged with the CLAWS POS tagger using the CLAWS 7 tag set.

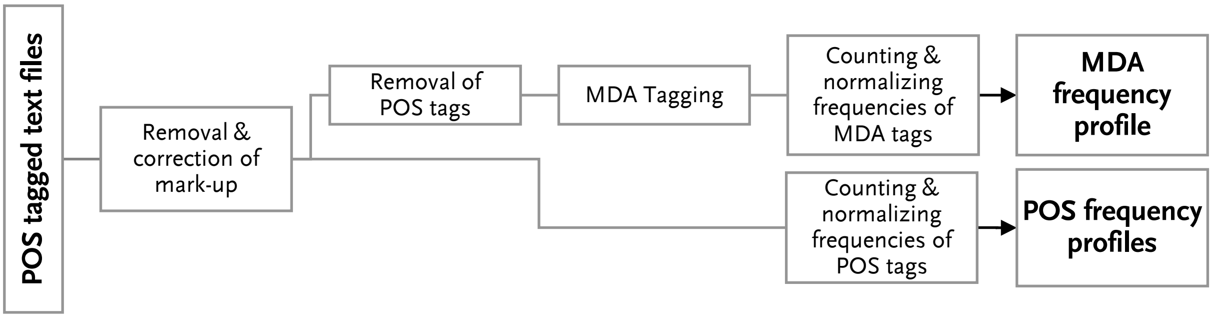

POS tagged text files
Removal & correction of mark-up
Removal of POS tags
MDA Tagging
Counting & normalizing frequencies of MDA tags
MDA frequency profile
Counting & normalizing frequencies of POS tags
POS frequency profiles

Figure 2: Compilation process of frequency profiles.

The first step in the data cleansing process is the removal of mark-up and extra corpus material, including extra-corpus speaker (speaker Z), indigenous (<indig>) and foreign words (<foreign>), editorial comments (<&>) and untranscribed text (<O>). Where mark-up indicated normalizations (e.g. for spelling corrections or indicating repetitions in spoken texts), the normalized text was used to increase the accuracy of the word counts and the tagging process. In ICE-CAN, the normalization of *gon_VVGK na_TO* to *going_VVG to_TO* was ignored to achieve consistency with other ICE components where the more informal and grammaticalized variant *gonna* was also not normalized. As the programming language *python* and regular expressions were used for the data cleansing, it was necessary to manually correct errors in the mark-up. Such errors mostly consisted of superfluous spaces (e.g. < /X>), unmatched opening or closing tags (e.g.: the use of <X> without the corresponding </X>), wrong placement and the resulting overlap of mark-up and the correction of utterance IDs (e.g. for text ICE-HK:W2B-023#1:1). The resulting files were compared to the original files using the freely available software *WinMerge* to verify the accuracy of the utilized regular expressions.

[24] I owe gratitude to Sebastian Pabel, who originally wrote the script for converting ICE-GB into a plain text and a POS tagged version.

In the next step, another version of the files without POS tags was created as the Multidimensional Analysis Tagger (MAT) (Nini 2019) requires plain text files as input. The MAT annotates the texts with 67 linguistically motivated tags (a description of which can be found in the appendix), which Biber (1988) used in his seminal work on distributional co-occurrence patterns of linguistic features in spoken and written registers. These tags are referred to as multidimensional analysis (MDA) tags. Tags include a wide variety of linguistic features, ranging from certain lexical classes (e.g. downtowners, hedges, emphatics), tense and aspect markers, subordination features (e.g. that adjective complements, WH-clauses) to reduced forms and dispreferred constructions (e.g. contractions, split auxiliaries). A full list of the features as well as the regular expressions used for the tagging process can be found in Nini (2014). A list of tags can also be found in *ICEtree*. Besides the individually tagged texts, part of the output of the MAT is also a file called *statistics.dat* that contains the frequencies of all tags per 100 words for each text.

The compilation of the frequency profiles underlying *ICEtree* is a relatively straightforward process. For the POS profiles, the pre-processed POS tagged files were converted to a single vertical file where each line holds one token, the corresponding POS tag and information on the utterance ID and the register. This representation was then used to aggregate the POS frequencies per text file using the programming language *R* (R Core Team 2019). Similarly, the frequency counts of the MDA tags as included in the output of the MAT were also imported into *R*. For easier interpretability, all frequency counts were converted to normalized frequencies (per 1,000 words, ptw). In order to alleviate distorting effects of a long-tailed frequency distribution of the individual tags, the frequencies were log transformed using the *R* function *log1p()*. Each text file is finally represented by a vector of logged frequencies of POS and MDA tags, respectively. The vector space representation of texts is illustrated in Table 11 below.

Variety	**Text ID**	**AT**	**CC**	**NN1**	**RR**	**PPIS1**	**PPY**	**...**
CAN	S1A-001	3.36	3.44	4.39	3.91	3.96	3.47	...
CAN	S1A-002	3.46	3.77	4.49	3.94	3.92	3.36	...
CAN	S1A-003	3.50	3.75	4.45	3.76	3.88	3.56	...
...	...	...	...	...	...	...	...	...

Table 11: Schematic illustration of a POS-based vector space representation of texts. Numbers represent logged frequencies per 1,000 words of the respective POS tag. AT= article, CC = coordinating conjunction, NN1 = singular common noun, RR = general adverb, PPIS1 = I, PPY = you.

3.2. Statistical methods employed

ICEtree allows the user to easily apply a range of statistical methods to the data. Although not all of the available methods are used in the case studies in chapter 5, their functions and possible applications are outlined in the description of *ICEtree* (see chapter 4). The focus of this section is therefore not the linguistic applications, but an outline of the methods and their interpretation. These outlines are kept deliberately brief for two reasons: i) their statistical properties and mechanics are described exhaustively in reference works on clustering (e.g. Everitt 2011, Moisl 2015 and sources quoted therein); and, maybe even more crucially, ii) application of statistical methods for linguistic enquiry requires a profound understanding of the methods *before* these are applied. The target audience of *ICEtree* is the statistically and corpus-linguistically versed linguist.

In an analogy to the user interface of *ICEtree*, the methods are grouped into methods for visualizing the dissimilarity or distances between profiles and methods for visualizing the frequencies of individual variables or tags. As *ICEtree* allows the aggregation of frequency profiles on several levels of granularity, the term *(frequency) profile* need not refer to profiles of individual texts, but can also refer to profiles of entire registers. If aggregation occurs, the aggregated profile is represented by a vector consisting of the mean values of the normalized frequencies of the selected tags.

3.2.1. Distance-based methods

The profiles, be it the POS profiles or the MDA profiles, are high-dimensional vector representations of the corpus texts. As such, these data cannot be visualized in a low-dimensional space. Commonly applied solutions to this problem are the reduction of dimensionality of the data or to the application of clustering. Where the goal of dimensionality reduction is to find a lower dimensional representation of the data while preserving as much of the variability inherent in the data as possible, clustering seeks to identify and visualize structure in the data. Although a wealth of methods is available (Moisl 2015 provides a detailed account of the application and mechanics of clustering and dimensionality reduction methods in linguistics), the methods I implemented in *ICEtree* are those that can be most commonly encountered in corpus-linguistic studies (see Moisl 2015: 277–301 for an overview of the use of cluster analysis in various linguistic disciplines). The implemented methods are agglomerative Hierarchical Cluster Analysis (HCA), unrooted phylogenetic trees[25], and metric Multidimensional Scaling (MDS).[26] While all of the above are based on distance matrices, the focus of the individual methods differ, and they each provide a unique perspective on the data. The figures in this section illustrate the different visualization techniques for the same data set (i.e. aggregated POS profiles of ICE-CAN). The example of HCA in Figure 3 encourages the interpretation of the data in terms of clusters.

[25] Although more and more studies seem to prefer phylogenetic networks using the NeighborNet Bryant & Moulton (2002) algorithm over unrooted phylogenetic trees using the neighbour-joining (NJ) algorithm Saitou & Nei (1987), at the date of writing the implementation of NeighborNet in *R* was still marked as experimental and thus excluded from *ICEtree* for practical reasons.

[26] MDS and MDA must not be confused. Where the former is a method for dimensionality reduction and projection of a high-dimensional space into a low-dimensional space, the latter makes reference to Biber's (e.g. (1988)) studies where he annotated texts with a large number of linguistic variables and applied a statistical method called factor analysis to identify co-occurrence patterns of these features.

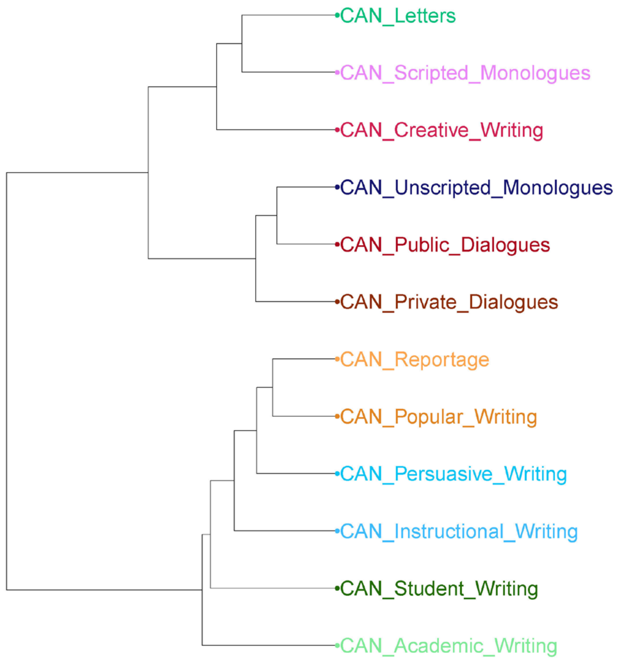


Figure 3: HCA (method:"ward.D2") of aggregated POS profiles of ICE-CAN.

Three clusters are visible: the first at the top comprises two more informal written registers (*letters* and *creative writing*) and one formal spoken register (*scripted speeches*) forming one cluster; the cluster in the middle contains the remaining, more spontaneous spoken registers (*unscripted speeches*, *private dialogues* and *public dialogues*); and the cluster at the bottom contains the remaining written registers. These clear-cut clusters are not visible in the unrooted phylogenetic tree in Figure 4. The intuitive interpretation follows the main branch or the "trunk" of the tree. *Private dialogues* and *academic writing* are the most distant registers, with a cline from informal spoken to formal spoken registers and formal written to informal written registers. In contrast, the MDS plot in Figure 5 does not guide the interpretation in such ways. The description of the individual methods below focuses on the technical aspects, whereas the potential linguistic applications are elaborated on in chapter 4 and exemplified in chapter 5.

What all methods described in this section have in common is that they take a distance matrix as input – i.e. a matrix containing pair-wise distances between the profiles (Table 1). These distances express the dissimilarity between the profiles: the larger the value, the more dissimilar the profiles are in terms of the logged frequency counts of tags. As the individual variables in the profiles are frequency data, Euclidean distance is used as the dissimilarity measure. The *R* function *dist()* is used for the calculation.

	CAN_Private	CAN_Public	CAN_Unscripted	CAN_Scripted
CAN_Private	0	2.56	3.47	6.28
CAN_Public	2.56	0	2.40	4.75
CAN_Unscripted	3.47	2.40	0	4.45
CAN_Scripted	6.28	4.75	4.45	0

Table 12: Exemplary distance matrix. Numbers indicate Euclidean distances between aggregated POS profiles of the included registers of ICE-CAN.

The visual output of HCA and unrooted phylogenetic trees are both referred to as *dendrograms*. A dendrogram consists of data points or leaf nodes (or leaves, in analogy to the tree metaphor), which represent individual data points. These leaf nodes are connected by branches. The point where branches connect are called splits or split nodes. The length of the branches visualizes the distances between leaves: the longer the branch, the more dissimilar the leaf nodes or split nodes are. Within hierarchical clustering, it can be further distinguished between agglomerative and divisive clustering (cf. Moisl 2015: 213–214). Both methods construct hierarchical dendrograms, but differ in where the construction process starts. Divisive (or top-down) clustering starts with a single cluster that contains all nodes (i.e. profiles) at the top and subdivides this cluster in consecutive steps until no further subdivision is possible. Agglomerative or bottom-up methods merge closely related data points into clusters, which in turn are merged consecutively until all data points are part of a single cluster. As agglomerative methods are by far the more common ones, only these are implemented in *ICEtree*. The main difference between the methods for agglomerative HCA is how exactly the distance between clusters is determined. While it is easy to identify two data points with the lowest dissimilarity index, doing the same for clusters is a more complex issue, and various methods are available and implemented in *R* (e.g. single linkage, complete linkage, Ward's method, etc.). Moisl (2015: 156–224) offers a detailed account of various clustering methods and the methods for cluster construction in HCA.

For the construction of unrooted phylogenetic trees, the neighbour-joining (NJ) algorithm (Saitou & Nei 1987) is used. NJ originated in bioinformatics and was developed to construct phylogenetic trees that help to visualize evolutionary distance and is just one of the many algorithms available. The reason for including this specific algorithm is that it i) is distance-based, hence does not require genetic sequences as input, ii) is fast in the construction of the trees – a requirement for interactive applications such as *ICEtree* – and iii) produces unrooted trees, which do not infer a common ancestor of connected leaf nodes. Huson & Bryant (2006) provide an overview of the various methods for constructing phylogenetic trees and their application in evolutionary studies. The construction process of the NJ-based unrooted phylogenetic trees is similar to that of agglomerative HCA in that it is a bottom-up, iterative process that is based on the distances between data points. One of the key differences in the selection of which data points to merge is that the NJ algorithm takes the sum of distances from one data point to all other data points into account. The first split is thus created between the two closest data points that also share the smallest sum of distances to all other data points. The tree creation process is often depicted as starting from a star-like dendrogram where all data points are connected and subsequent splits are inserted based on the distance criterion of NJ (cf. Saitou & Nei 1987: 408–411). Although the branch length was originally meant to visualize the evolutionary distance and speed of development, it effectively visualizes relatedness between data points and need not be interpreted from an evolutionary perspective. Although similar to HCA, the output of NJ is an unrooted tree and as such the focus is on relatedness between data points and clusters, rather than on hierarchical structures (as is the case for HCA).

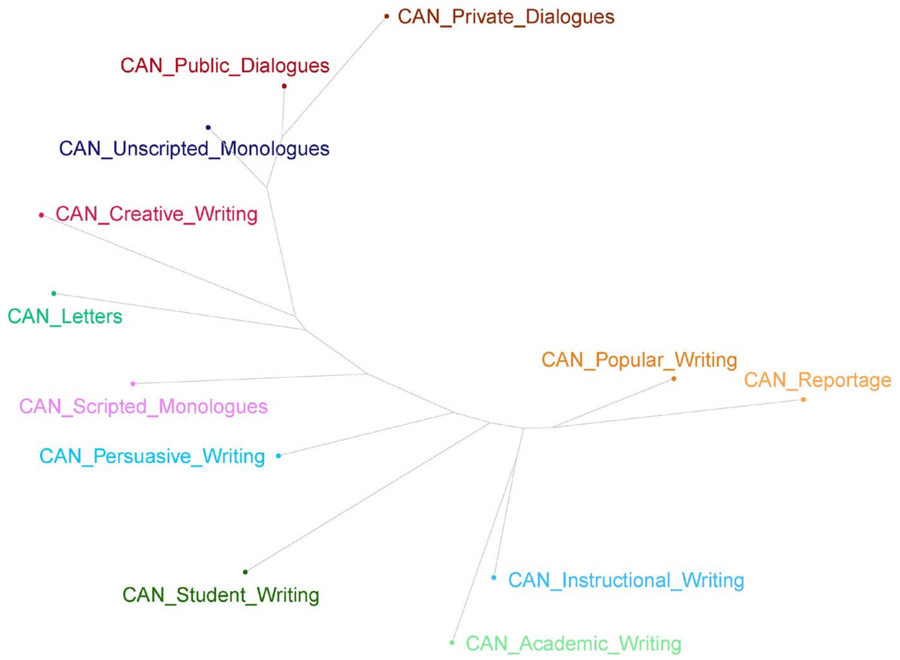


Figure 4: Unrooted phylogenetic tree of aggregated POS profiles of ICE-CAN.

Multidimensional Scaling (MDS) differs from the previous two methods in that it does not group data points by some distance criterion, but instead visualizes high-dimensional data in a low-dimensional (often 2D) space. MDS is well documented (e.g. Cox & Cox 2001, Borg & Groenen 2005 or Borg et al. 2013) and used extensively in a wide range of scientific disciplines. Although many variants of MDS exist (cf. Borg et al. 2013: 37–48), MDS is often classified based on the quality of the data: *metric* or *classical* variants of MDS are used for metric data, and *non-metric* MDS for categorical or ordinal data. As the profiles used in *ICEtree* are based on frequencies of features, classical MDS is used. In what follows, the term *MDS* is used as shorthand for classical MDS. While a description of the statistical process is omitted here (and is indeed not easily explained, see Borg et al. 2013: 81–84 for details), the underlying idea and the interpretation are relatively intuitive. The MDS algorithm takes a dissimilarity matrix with Euclidean distances as input and seeks a low-dimensional representation of the data while preserving as much variability of the data as possible. Typically, two-dimensional MDS solutions are used. The interpretation is also relatively straight-forward: the distance between two data points denotes (dis-)similarity. With regard to the present study, this means that the closer two data points appear in the resulting MDS solution, the more similar their frequency profiles are.

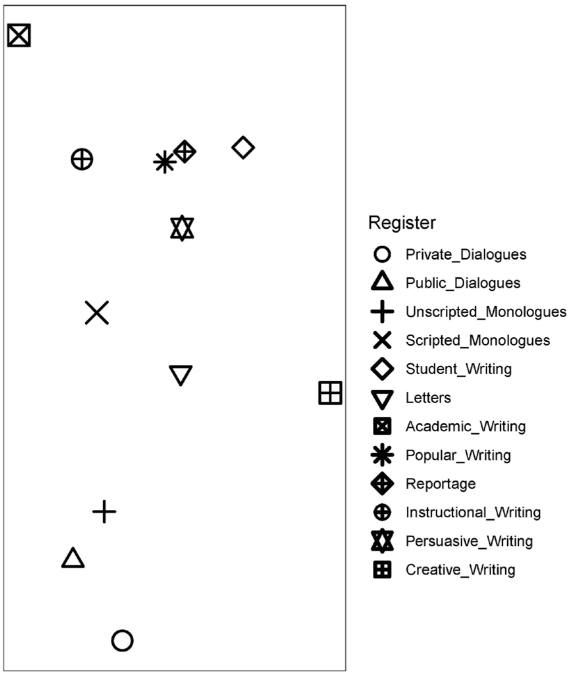


Figure 5: MDS of aggregated POS profiles of ICE-CAN.

3.2.2. Investigating frequencies of variables: mean frequency, standard deviation & random forests

The second major perspective on the data available in *ICEtree* visualizes the normalized frequencies of individual tags. While plotting the frequencies of all tags is possible, inspecting such a large number of tags is practically not feasible. For this reason, three methods for ordering tags are implemented in *ICEtree* and the user can choose to investigate the top *n* tags. The implemented methods are: Order tags i) by overall mean frequency, ii) by their standard deviation, and iii) by their importance ratings extracted from a random forest model. For i) and ii), the tags with the highest overall mean frequency or standard deviation of the frequencies are plotted at the top of the chart. While insightful in their own right, during my own analyses, it emerged that these are not the most helpful if distinct clusters are visible in the distance-based methods.

The intuitive question in such a scenario is which variables show the greatest difference between the clusters. For this purpose, the profiles and the information on variety and register are used to train a random forest model using the *R* package *randomForest* (Liaw & Wiener 2002). Although a random forest is a machine learning algorithm that is commonly used in to classify new data based on a set of training data, it can also be used for feature selection – i.e. to discover variables that are most informative for dividing data points into pre-defined groups. The application of a machine learning algorithm can be essentially broken down into three phases. In the first phase, the algorithm is provided with training data where the target classification of the data points is known (hence training phase). In this phase, the algorithm identifies the variables and (threshold) values of such that are most distinct to the groups present in the training data set. In the second phase, a second data set where the classification is also known is used to test the precision of the classification and manually modify the classification parameters if necessary. In the third phase, the algorithm is used to classify new data where the classification is not known in advance. The subsequent outline of the applied method is based on Aggarwal (2018: 142–147), who provides an introduction to random forests and decision trees. Since a random forest is an ensemble method that is based on decision trees, the latter are described first. A decision tree is a machine learning algorithm that partitions data in a top-down, tree-like fashion along so-called splits. Splits can be understood as conditional statements relating to the value of a variable. Such a conditional statement can be a threshold value or a range of values of the variable that are characteristic to groups of data points supplied to the algorithm during the training phase.

The growing of a decision tree is best explained with a practical example. Suppose the goal is to use a decision tree to classify texts into involved or highly informational texts based on the linguistic features Biber (1988) identified in his multidimensional analysis of registers. The data set used in the training phase would then consist of a set of texts where the classification (involved vs. informational) is known in advance. The data are represented by a matrix where each text is expressed as a vector of frequencies of the linguistic features. The algorithm identifies those features that draw the starkest contrast between the two pre-defined groups of texts in the training data and determines threshold frequencies for the subsequent classification of unknown texts. Based on the factor loadings presented in Biber (1988: 102), we would expect the algorithm to identify the frequencies of private verbs, *that*-deletions, contractions, present tense verbs and 2nd person pronouns (among others) as the features most relevant for the classification. The algorithm stops once all branches of all splits consist of data points of a single group. As an example (Figure 6), I trained a decision tree for all texts in the registers *private dialogues* and *academic writing* in ICE-CAN. As variables I used the normalized frequencies (per 1,000 words) of contractions, *that*-deletions and private verbs. As all texts can be successfully classified based on the frequency of contractions, the algorithm did not include all variables. As threshold value, the algorithm determined a frequency of 15.85 contractions per 1,000 words. Any texts that exhibit less contractions per 1,000 words are classified as academic writing, and all other texts as private dialogues. The R package *tree* (Ripley 2019) was used for training the decision tree.

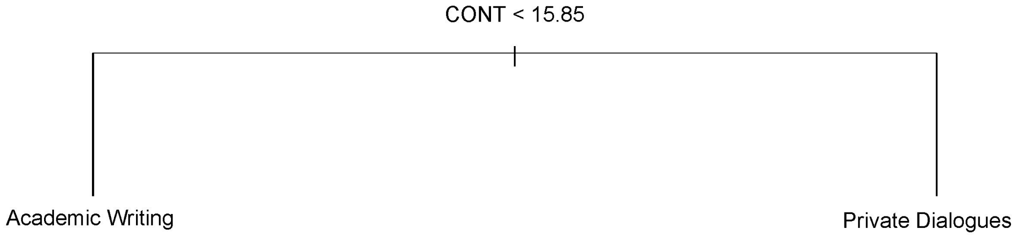


Figure 6: Decision tree for all MDS profiles of all texts in *academic writing* and *private dialogues* in ICE-CAN.

To overcome the shortcomings of decision trees (overfitting among other things), the random forest algorithm trains *n* decision trees for each split and chooses the best feature out of a randomly selected subset of *r* features. After the training process, the final tree is derived by averaging all previously grown randomized trees. Aggarwal (2018: 146) concludes that this process "results in more robust predictions" compared to decision trees. In general, he states that random forests "are known to be highly robust and accurate" (2018: 142). Although random forests are primarily used as classifiers or regressors in a machine learning context, they can also be used for what is called 'feature extraction' since importance measures for each feature can easily be calculated or are already part of the output of the respective implementation. So instead of using the random forest model to classify new data, the variables which the model determined as most informative for the classification process in the training phase are extracted. The advantage of utilizing random forests compared to other methods in this scenario is that they allow the identification of individual features instead of groups of features. This is especially useful for register comparisons on a finer level of granularity, where only a small set of features should vary.

The variables can be extracted from the model using the *R* function *importance()*. It is only the order of the variables as determined by the model that is relevant for the subsequent plotting of the frequencies of variables. These frequencies are visualized in form of a dot plot (see Figure 7). Depending on the number of iterations set for the training process, the exact order of the variables might vary slightly. A higher number of iterations results in more stable importance ratings and thus in a more stable order of the variables. I must again caution against the over-interpretation of quantitative results. The order of the variables and the differences of the frequencies of these variables should be understood as a point of departure for qualitative analyses – a method that will be adopted in the analytical chapters to follow.

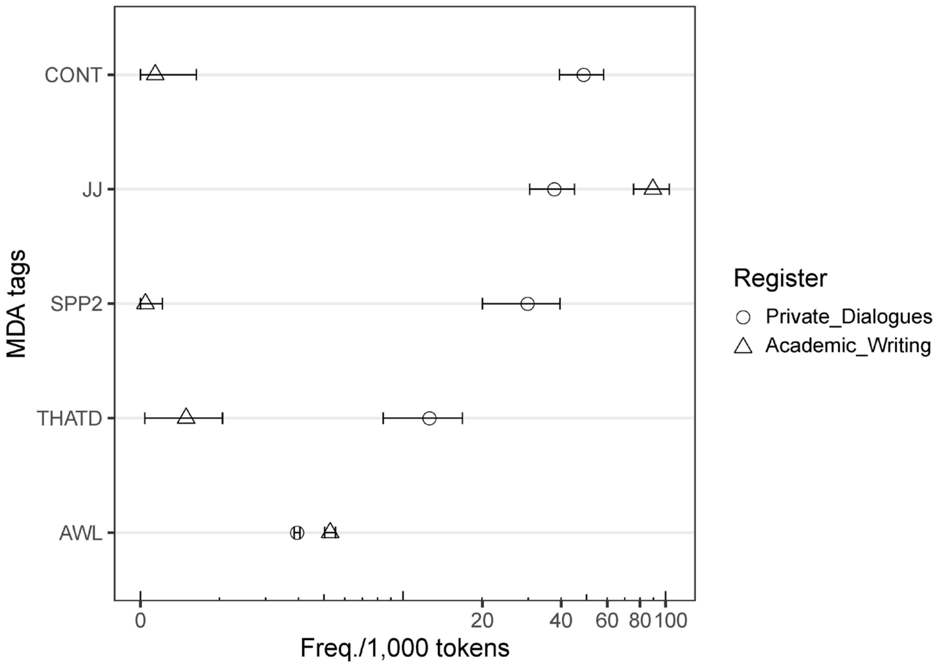


Figure 7: Average frequencies of the MDA tags for *private conversations* and *academic writing* in ICE-CAN with the highest importance ratings (error bars indicate standard deviation). CONT = contractions, JJ = general adjective, SPP2 = second person pronouns, THATD = that deletion, AWL = average word length.

4. *ICEtree*: Interactive data visualizations of ICE

At its core, *ICEtree* is an interactive web application (also referred to as *web app* or simply *app*) for exploring components of ICE from a quantitative perspective. The app allows the user to apply a range of methods for visually exploring similarities between (groups of) texts on various levels of granularity. Although a fundamental understanding of the statistical methods employed is required, one of its key selling points is certainly that no knowledge of programming is necessary to apply these methods. It therefore extends the corpus linguists toolbox of established methods such as concordancing, (key) word lists, *n*-grams and collocation analysis by methods for clustering and dimensionality reduction. While clustering (both exploratory and confirmatory) have been used extensively in corpus linguistics, the novelty of *ICEtree* is that it offers an intuitive interface to clustering texts and allows the user to easily identify which variables set observed clusters apart.

ICEtree is written in the programming language *R* and utilizes a range of so-called libraries, mainly for achieving its interactivity, such as *shiny* (Chang et al. 2019) or *plotly* (Sievert 2020), increasing speed, such as *data.table* (Dowle & Srinivasan 2019), and applying statistical techniques, e.g. *randomForest* (Liaw & Wiener 2002).[27] Both the source code and a standalone are available in a repository on the open science framework (OSF) platform: https://osf.io/ztfsx/.

The simplest way of launching *ICEtree* is to start the standalone version by double-clicking the "START_ICEtree.bat" file in the main folder of the standalone version. The advantage of using the standalone version is that *R* and all required packages in the correct versions are already bundled together. The standalone version utilizes a modified skeleton for standalone shiny apps as provided by Pang (2019). The script "START_ICEtree.bat" opens the bundled version of R in the background and starts the app. Once the app is running, the script opens a web browser in which the app then is displayed. Running *ICEtree* from its source requires installing *R* and all packages required manually. More information on running *shiny*-based apps from RStudio (RStudio Team 2019) can be found at https://shiny.rstudio.com/.

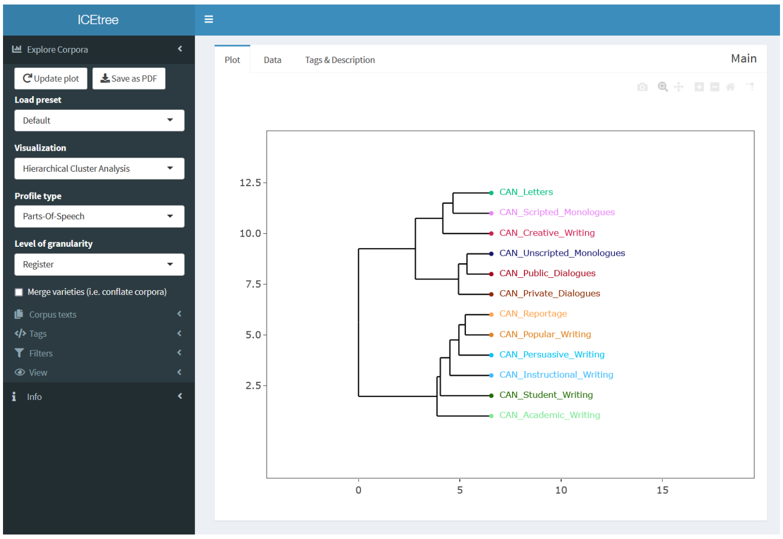
The first section of this chapter describes the program *ICEtree* and presents its functions in details with a focus on potential linguistic applications of the individual functions. The description of the functions also serves as a model of a potential workflow where *ICEtree* is used to zoom in on patterns that already emerge on a coarse level of granularity. In the course of this description, cautious interpretations of the output of these functions are also offered and possible directions for future development are hinted at. As the application is under constant development, the most recent version of the description of all available options will be made available in the OSF repository. The focus here lies on the main functions and their application. Additionally, the last

[27] A full list of utilized packages and the respective references are provided in the documentation in the app itself.

section of this chapter reflects on developing applications for linguists and discusses dissemination strategies.

4.1. *ICEtree:* User interface and functions

Once started, the interface of *ICEtree* is shown in a web browser. It is essentially divided into two parts: the options and navigational panel on the left and the main or plot panel on the right (Figure 8). The plot panel contains three tabs: "Plot", "Data" and "Tags & Description". The first tab is where the rendered plot is drawn, the second provides the data the plot is based on in a tabular format and the third offers a list of the tags and their description for the chosen profile type. The options panel contains two collapsible menus: **Info** and **Explore Corpora**. The **Info** menu contains information on the app and the documentation of the individiual functions of the app.



Load preset: Default

Visualization: Hierarchical Cluster Analysis

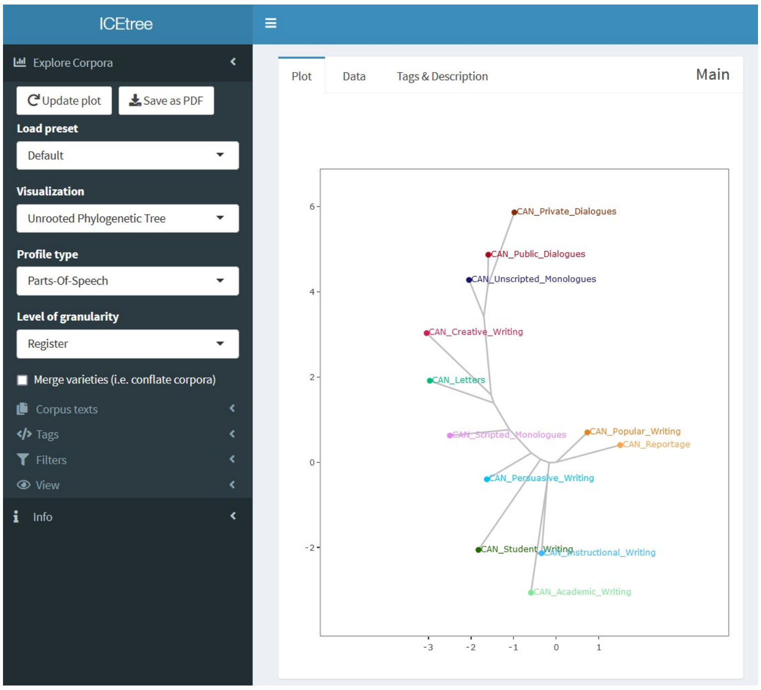
Profile type: Parts-Of-Speech

Level of granularity: Register

□ Merge varieties (i.e. conflate corpora)

Figure 8: Main interface of *ICEtree* with the options and navigational panel (left) and the main or plot panel (right).

The **Explore Corpora** menu forms the heart of the app (Figure 8). Here, the user is presented with a wealth of options.



Load preset: Default

Visualization: Unrooted Phylogenetic Tree

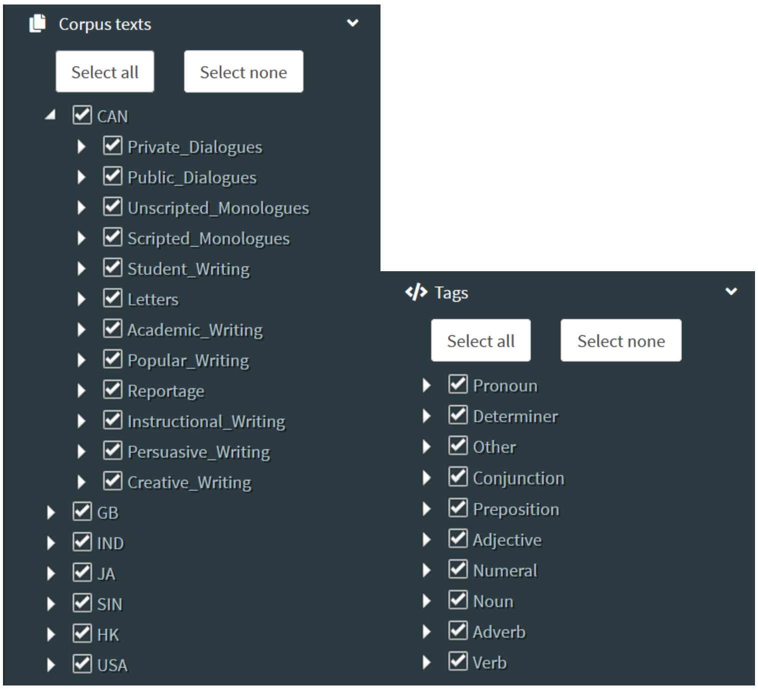
Profile type: Parts-Of-Speech

Level of granularity: Register

□ Merge varieties (i.e. conflate corpora)

Figure 9: “Explore corpora” panel in ICEtree. The plot on the right shows the unrooted phylogenetic tree of the aggregated POS profiles of the registers of ICE-CAN.

At the top are two buttons for updating and exporting the visualization of the data (**Update plot** & **Save as PDF**) and dropdown menus for loading presets (**Load preset**), the method of visualization (**Visualization**), the profile type (**Profile type**, i.e. whether the data points should be based on parts-of-speech or linguistically motivated tags), the level of granularity (**Level of granularity**, i.e. whether individual text profiles should be aggregated) as well as a checkbox that tells *ICEtree* to disregard information on variety and treat all texts as if they were from just one variety (**Merge varieties (i.e. conflate corpora)**). The submenu **Corpus texts** (Figure 10: left) allows the user to select individual texts, groups of texts or entire corpora. Unless a specific preset is loaded, all texts of ICE-CAN are selected per default. The selectable corpus texts include the corpus extension and the reclassification described in section 3.1.1.2. In the submenu **Tags** (Figure 10: right), the user can decide which tags of the respective profiles should be included in the computation.

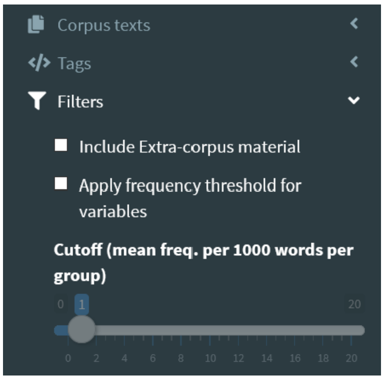


- ☑ CAN
  - ☑ Private_Dialogues
  - ☑ Public_Dialogues
  - ☑ Unscripted_Monologues
  - ☑ Scripted_Monologues
  - ☑ Student_Writing
  - ☑ Letters
  - ☑ Academic_Writing
  - ☑ Popular_Writing
  - ☑ Reportage
  - ☑ Instructional_Writing
  - ☑ Persuasive_Writing
  - ☑ Creative_Writing
- ☑ GB
- ☑ IND
- ☑ JA
- ☑ SIN
- ☑ HK
- ☑ USA

- ☑ Pronoun
- ☑ Determiner
- ☑ Other
- ☑ Conjunction
- ☑ Preposition
- ☑ Adjective
- ☑ Numeral
- ☑ Noun
- ☑ Adverb
- ☑ Verb

Figure 10: Left: Submenu "Corpus texts" with all texts of ICE-CAN selected. Right: Submenu “Tags” with all POS tags selected.

In **Filters**, the user can set a frequency threshold for features and decide whether extra-corpus material should be included. The latter option can be useful to investigate which of the selected texts contain the greatest amount of extra-corpus material and whether the profiles of those texts behave differently. Extra-corpus material in spoken conversations often indicates non-corpus speakers and may occur in situations where a researcher talked to a native speaker of the variety under investigation.



□ Include Extra-corpus material

□ Apply frequency threshold for variables

Cutoff (mean freq. per 1000 words per group)

Figure 11: Submenu "Filters".

Figure 12 shows the MDS plot of the POS profiles of all texts of the register private conversations in ICE-SIN. The amount of extra-corpus material is given in percent and indicated by a red hue. As the scale is determined automatically, bright red indicates that the text includes around 30% extra-corpus material (e.g. ICE-SIN:S1A-090, the bright red outlier at the bottom of the plot, consists of 34% extra-corpus material). In this case, quite a number of texts include extra-corpus material of around 15%. Yet, no clear patterns is visible, which indicates that these profiles exhibit no systematic difference from texts that include little to no extra-corpus material – at least with regard to their POS profiles.

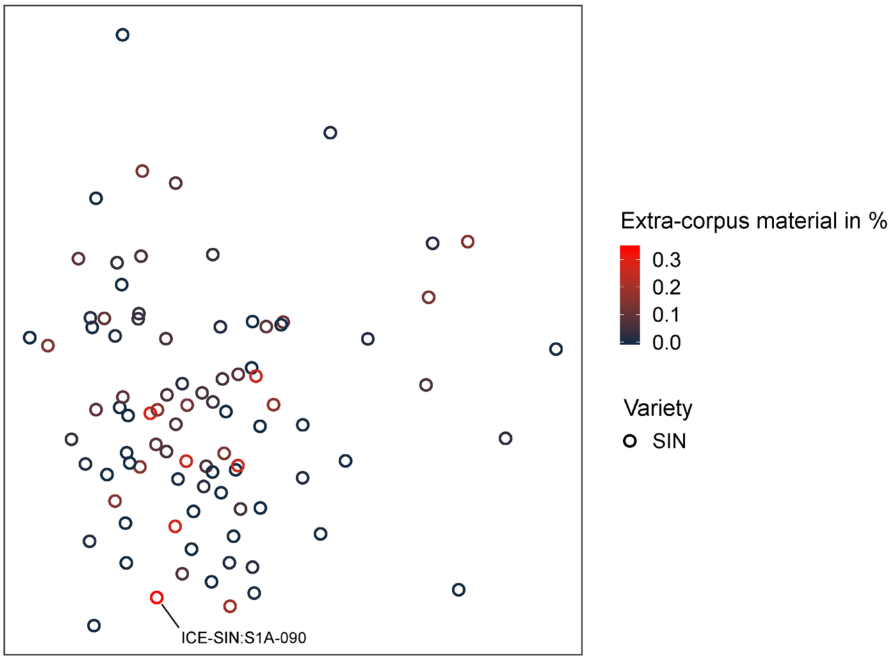
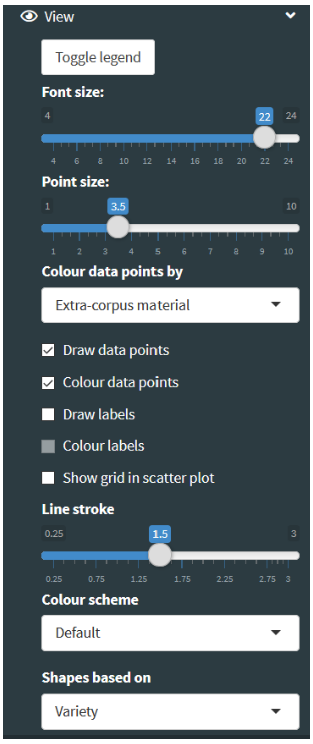


Figure 12: MDS plot of all POS profiles of private conversations in ICE-SIN, amount of extra-corpus material indicated in red.

The options in **View** determine the exact gestalt of the visualization (Figure 13). The available options depend on the previously selected settings (especially the selected visualization technique and the level of granularity). Settings present for all visualization techniques are **Font size**, **Colour data points by** and **Colour scheme**. The **Font size** slider allows the user to set the relative font size of labels drawn within the plot. **Colour data points by** indicates what determines the colour of the data points, which can be either the variety, a level of granularity or the amount of extra-corpus material. The options available depend again on other settings (e.g. level of granularity and visualization technique). For **Colour scheme**, the user is presented with two choices: “Default” or “Rainbow”. The default colour scheme consists of maximally distinct colours, whereas for the colour scheme “Rainbow” the individual registers are positioned on a rainbow-coloured gradient in the order as they appear in the sampling scheme.



View

Toggle legend

Font size: 22

Point size: 3.5

Colour data points by: Extra-corpus material

☑ Draw data points

☑ Colour data points

□ Draw labels

□ Colour labels

□ Show grid in scatter plot

Line stroke: 1.5

Colour scheme: Default

Shapes based on: Variety

Figure 13: Submenu "View".

In what follows, I will describe and illustrate the individual visualization techniques and their options. The dropdown list **Visualization** contains five options of which the first three are grouped into "Distances between profiles". The methods in "Distances between profiles" are all visualizations of distance matrices and available methods are: "Multidimensional Scaling", "Hierarchical Cluster Analysis" and "Unrooted Phylogenetic Tree". The two remaining methods are "Normalized frequencies" and the experimental "Average distance between profiles". As the name suggests, "Normalized frequencies" visualizes the normalized (per 1,000 words) frequencies of the selected tags in the selected data. "Average distance between profiles" shows the average Euclidean distance between (groups of) profiles in a dot plot.

4.1.1. Visualization: Distances between profiles

When **Visualization** is set to one of the methods in the group "Distances between profiles", *ICEtree* performs the following steps in sequence: i) aggregation of the profiles on the selected level of granularity, ii) calculation of Euclidean distances between the selected profiles, iii) application of either metric MDS, HCA or NJ, and iv) plotting the data.

When set to "Multidimensional Scaling", the output is a scatter plot where each symbol represents either the profile of a single text or of a group of texts. If all options are set to their default values, *ICEtree* aggregates all profiles of ICE-CAN on a register level and colour-codes the data points by register. Figure 14 was created using the default settings, only "Shapes based on" was set to "Off" and the point and font size were increased. Each data point represents an aggregated POS profile of ICE-CAN and colour indicates the register. The interpretation is relatively straightforward: proximity of two data points translates as similarity between the profiles.

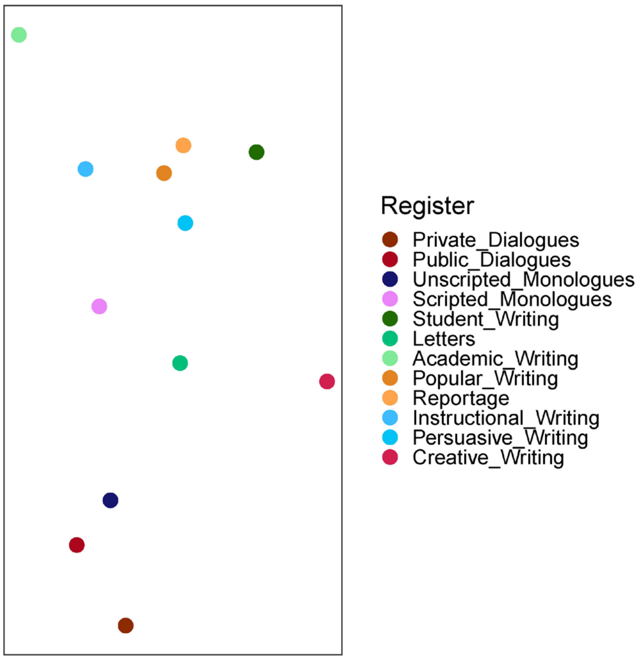
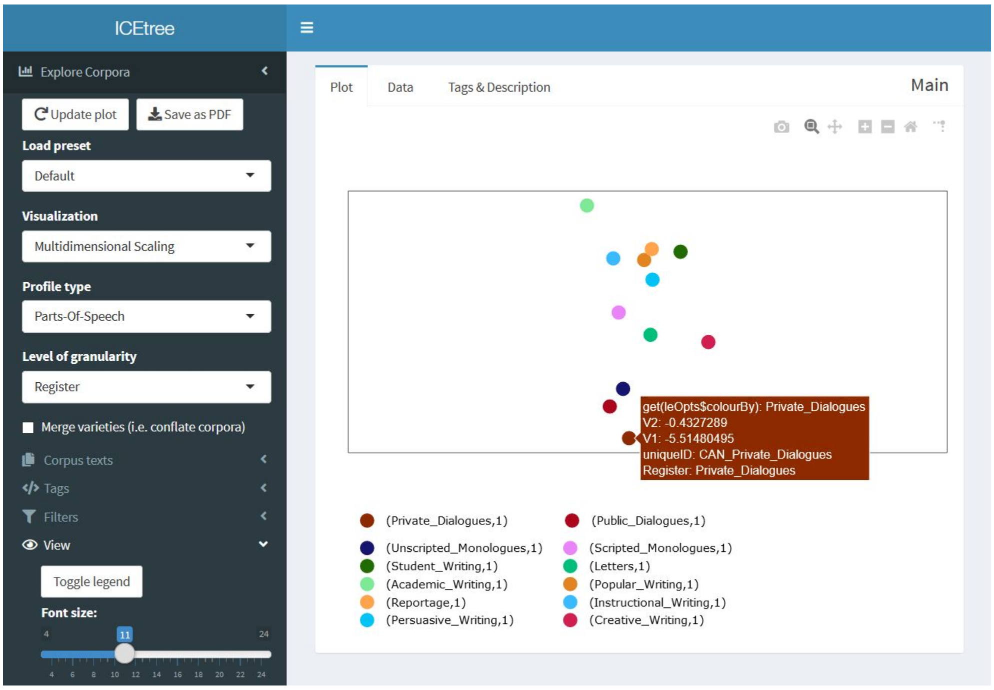


Figure 14: MDS of aggregated POS profiles of all registers in ICE-CAN.

In contrast to factor analysis, the axes in MDS should not be overinterpreted (see Borg et al. 2013: 70–74 on this subject) and as a result the axis labels are omitted in the plot. Since the interactive plots are rendered using the *R* library *plotly*, the user is provided with additional information when hovering over data points (see Figure 15) and can show and hide data points by clicking on groups in the legend. Additionally, zooming is also possible.



Load preset: Default

Visualization: Multidimensional Scaling

Profile type: Parts-Of-Speech

Level of granularity: Register

□ Merge varieties (i.e. conflate corpora)

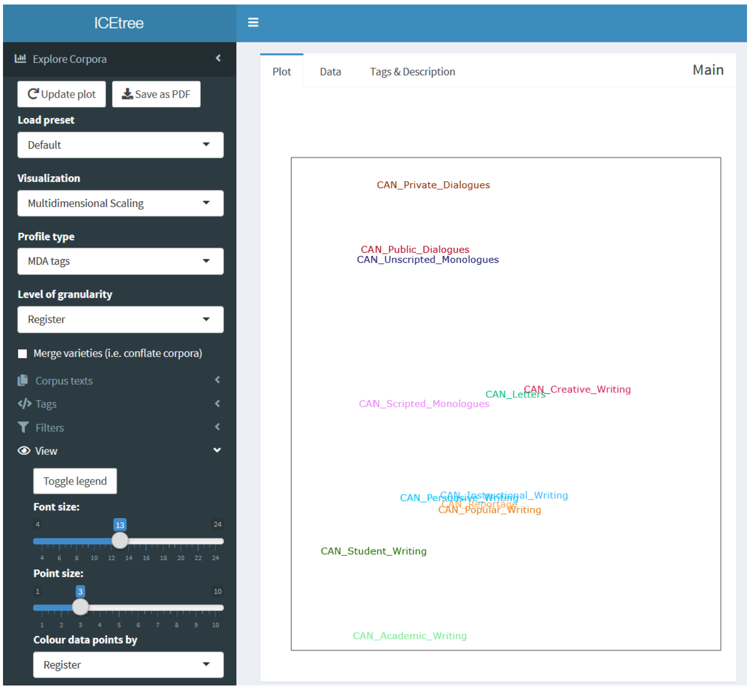
Font size: 11

```
get(leOpts$colourBy): Private_Dialogues
V2: -0.4327289
V1: -5.51480495
uniqueID: CAN_Private_Dialogues
Register: Private_Dialogues
```

Figure 15: Interactive MDS plot with mouse over function illustrated.

Upon closer inspection of Figure 14, an interesting pattern already emerges. Where the data points of spoken registers form a cluster towards the bottom of the plot, the top half of the plot is populated by the written registers. What can further be observed is that within the spoken registers there is a gradient from informal registers at the bottom, to more formal ones towards the middle of the plot. A similar pattern is visible for the written registers although the gradient is orthogonal to that of the spoken registers with formal registers (academic writing, instruction writing) populating the left and less formal written registers (letters, creative writing) towards the right of the plot.

This pattern is even more pronounced if **Profile type** is set to “MDA tags” (Figure 16). In this plot, the options in the submenu **View** were used to draw coloured labels instead of symbols. The center of the label indicates the position of the data point. Although both profile types offer very similar results here, Vetter (to appear) compares and evaluates the outcome for both in his analysis of the register *press editorials*.



Load preset: Default

Visualization: Multidimensional Scaling

Profile type: MDA tags

Level of granularity: Register

□ Merge varieties (i.e. conflate corpora)

Font size: 13

Point size: 3

Colour data points by: Register

Figure 16: Interactive MDS plot, but based on MDA tags.

The submenu **Corpus texts** is organized hierarchically, each group label is clickable and allows for easy selection of registers or entire varieties. For the next MDS plot (Figure 17), all texts of five ICE varieties (ICE-CAN, ICE-GB, ICE-IND, ICE-JA, ICE-SIN) were selected by clicking the checkbox in front of the abbreviated variety names. For this plot, the additonally sampled corpus material as described in section 3.1.1.2 was excluded. **Profile type** was again set to "Parts-Of-Speech" and, unless otherwise indicated, all plots in this section use this type of profile and data selection. The colour indicates register and the symbol indicates variety. In the newly created plot, the earlier described organization of registers still holds.

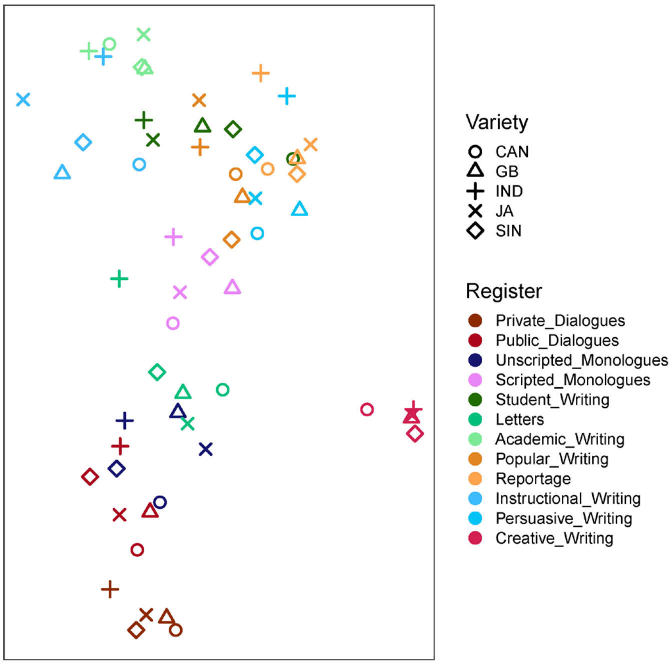


Figure 17: MDS plot of aggregated POS profiles of ICE-CAN, ICE-GB, ICE-IND, ICE-JA and ICE-SIN.

In Figure 18, **Visualization** was set to "Unrooted Phylogenetic Tree" with both **Draw labels** and **Colour labels** enabled. With the highly involved *private dialogues* and the highly informational *academic writing and instructional writing* forming the two poles of the dendrogram, the organization can be best described as two overlapping gradients where the spoken and written registers are aligned along a cline from involved to informational (see 5.3. for a discussion of this pattern in the context of the case studies presented in 5). The pattern bears a strong resemblance to the findings of Biber (1988), especially where the registers of the LOB and the London-Lund Corpus of Spoken English (LLC) are organized by their mean scores on the first dimension ("involved vs. informational production") of the factor analysis (cf. Biber 1988: 128). The difference between the Jamaican and Canadian and the remaining persuasive writing sections is more pronounced and more easily detectable in the dendrogram. Incidentally, it is this view that inspired the closer inspection of press editorials as presented here later. The dendrogram also hints at differences in the composition of student writing, with the respective sections ICE-SIN, ICE-GB and ICE-CAN forming a cluster more closely related to persuasive writing, and the sections of ICE-JA and ICE-IND more closely related to academic writing.

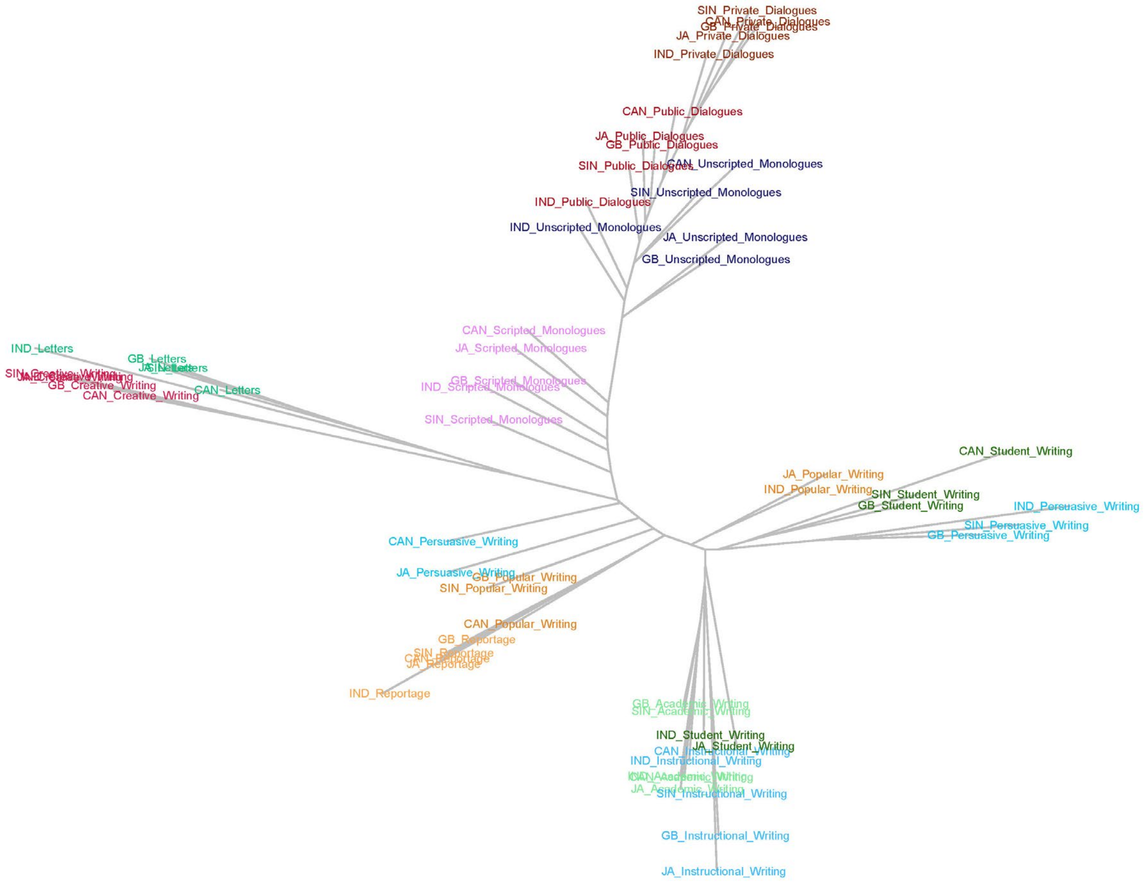


Figure 18: Unrooted phylogenetic tree based on aggregated POS profiles of ICE-CAN, ICE-GB, ICE-IND, ICE-JA and ICE-SIN.

When setting **Visualization** to “Hierarchical Cluster Analysis”, the distance matrix is not subjected to MDS or the NJ algorithm, but instead passed to the *R* function *hclust()*. Additionally, the user is presented with a new slider **Margin (increase if labels truncated)** with which the margin of the plot can be manually manipulated as labels of data points sometimes are truncated. The user can further choose the **Agglomeration method** to be used in the HCA. The function *hclust()* offers a range of methods for determining to which cluster a data point belongs and all available methods are selectable in *ICEtree*. Available options are “ward.D”, “ward.D2” (default), “single”, “complete”,”average”,”mcquitty”, “median” and ”centroid”. Moisl (2015: 201–224) provides an in-depth description of the most common methods. For the resulting plot (Figure 19), the checkboxes **Draw labels** and **Colour labels** were set, **Agglomeration method** was left at the default value (“ward.D2”) and the margin was slightly increased as some labels were truncated. By and large, and quite unsurprisingly, the results of HCA are compatible with those gleaned from the MDS and the unrooted phylogenetic tree. Except for *unscripted monologues* and *public dialogues*, the data points cluster relatively clearly by register. From the top, two main clusters are visible. One comprises more informational written registers and *scripted monologues*, the other the more involved written registers (*letters* and *creative writing*) and the more informal and spontaneous spoken registers (*private dialogues*, *public dialogues* and *unscripted monologues*). Both main clusters are again divided into two larger clusters. The predominantly written cluster is divided into a cluster containing the highly informational registers (i.e. *academic writing*, *instructional writing* and *student writing*) and another cluster containing news writing and the medially spoken but conceptually rather written (cf. Koch & Oesterreicher 2012) scripted monologues. *Popular writing* takes a middle ground and can be found in both clusters. What is noteworthy and, while visible, does not feature as prominently in the MDS plots is that the Canadian *student writing* appears to be more closely related to *persuasive writing* than to the other *student writing* sections of the remaining varieties. Although the HCA shows that the British and Singaporean *student writing* are closely related, their closer relatedness to *popular writing* than to *academic writing* cannot be directly seen in this plot – as compared to MDS and the unrooted phylogenetic tree.

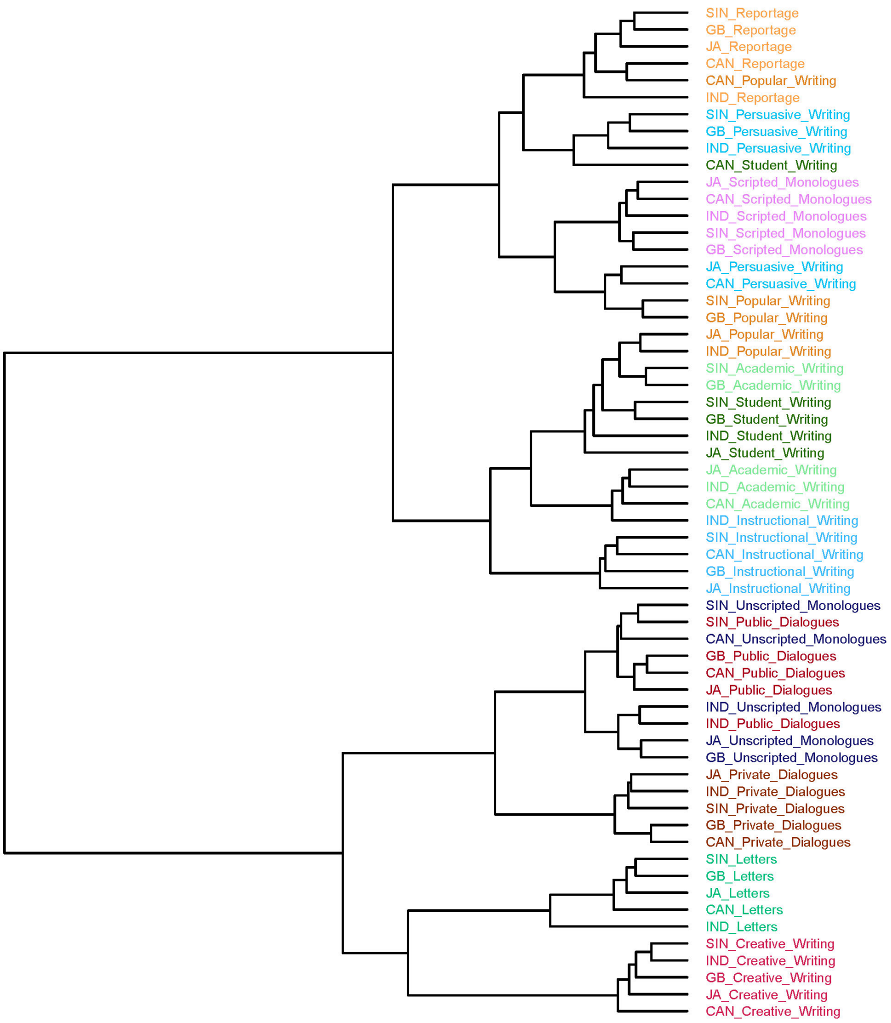


Figure 19: HCA of aggregated POS profiles of all registers in ICE-CAN, ICE-GB, ICE-IND, ICE-JA and ICE-SIN.

To further investigate the patterns observed for the student writing section, the data selection was limited to the texts of this section. During my own investigations, traversing down the levels of granularity and using *plotly*'s interactive functions for selectively hiding and showing data points turned out to be helpful for obtaining an understanding of the data. Where the left panel in Figure 20, an MDS plot on the topmost level of granularity, does not provide any meaningful insights, choosing "Subregister" in the **Level of granularity** dropdown menu (right panel) reveals that the two subregister in *student writing*, *untimed essays* and *exam scripts* or *timed essays*, cluster largely by variety.

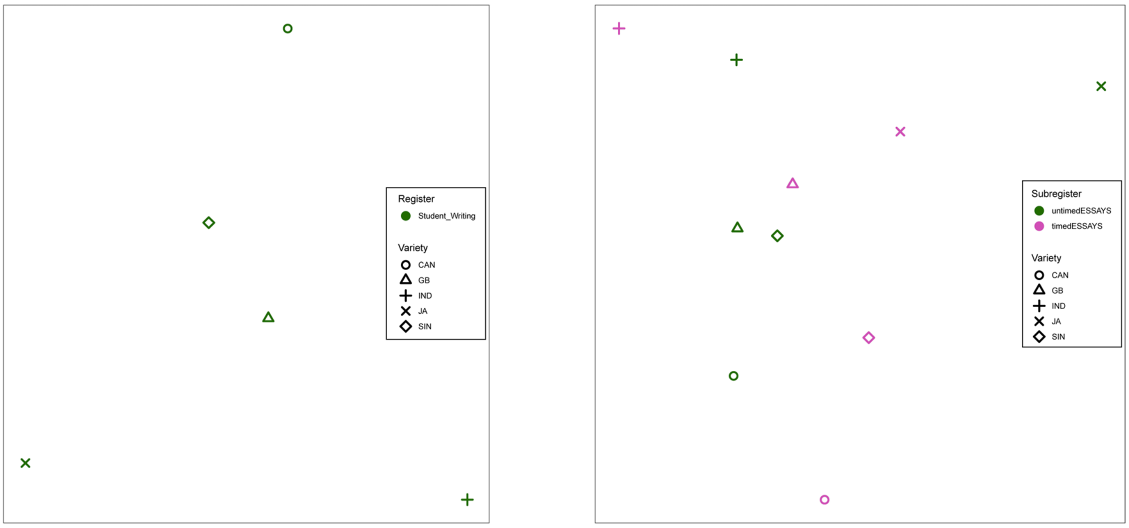


Figure 20: Left: MDS of aggregated POS profiles of *student writing* in ICE-CAN, ICE-GB, ICE-IND, ICE-JA and ICE-SIN. Level of granularity set to "Register". Right: Same, but level of granularity set to "Subregister".

When setting **Level of granularity** to "Texts", each data point represents the profile of an individual text file. Due to the great number of data points in Figure 21, the plot appears fairly crowded and identifying patterns is difficult.

However, using the interactive visualization in *ICEtree* allows for toggling the visibility of groups of data points. By successively hiding all but one group of data points, two subregisters that form uncharacteristically dense clusters can be identified: the Canadian *timed essays* and the Jamaican *untimed essays*. Possible reasons for this pattern are discussed in the next section. Figure 22 shows each layer of the MDS plot as individual panels.

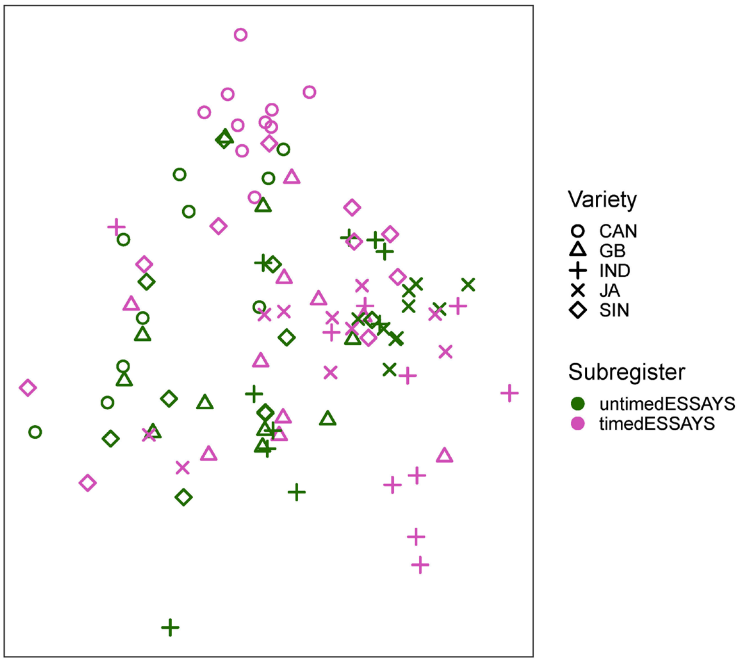


Figure 21: MDS of texts in *student writing* in ICE-CAN, ICE-GB, ICE-IND, ICE-JA and ICE-SIN. Level of granularity is set to "Texts".

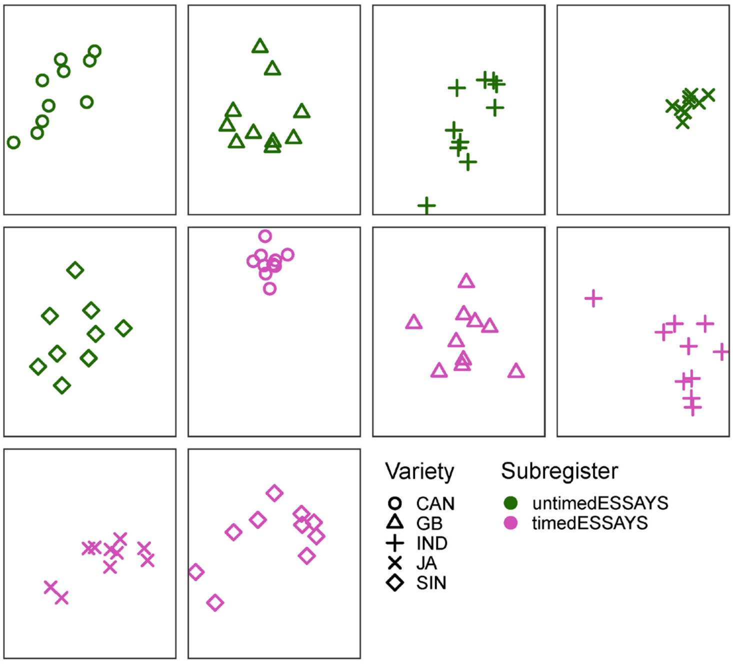


Figure 22: Each panel represents one layer of the MDS plot of the aggregated POS profiles of *student writing* with level of granularity set to "Texts".

4.1.2. Visualization: Average distance between profiles

Although not grouped with the distance-based methods introduced in 4.1.1, the visualization "Average distance between profiles" is also based on Euclidean distances. However, instead of applying some form of dimensionality reduction or clustering, this method calculates the average distance between text profiles of a group and produces a dot plot. Unless the option **Merge varieties (i.e. conflate corpora)** is set, all data points of one variety and one register or subregister (depending on the selected level of granularity) are considered a group. For this visualization technique, the user can additionally choose how the average distance is computed. Available options in the dropdown menu Measure are "Mean distance between profiles", where the means of all distances between all data points of a group are calculated, and "Distance from centroid", where the mean distance from all data points of a group to the group centroid is calculated. While both options yield similar results, "Distance from centroid" is more susceptible to outliers. Please note that this visualization technique is labelled experimental as its development and testing is, at the time of writing, still in the early stages and will be improved as the development of *ICEtree* progresses. However, from my own investigations it appears that outliers in this plot often hint at discrepancies in sampling between components of ICE. If the outlier is characterized by a lower or higher than average mean distance, the group of data points exhibits respectively less or more variation in terms of sampled (sub-)registers or topics than other groups. This is best illustrated by returning to the ICE register *student writing*. The mean distance between the text profiles of one subregister and variety for the register *student writing* for the Canadian, Singaporean, Indian, Canadian, British and Jamaican components of ICE is visualized in Figure 23. What is visible in the MDS plots in Figure 22, stands out even more clearly in Figure 23: the Canadian *timed* and the Jamaican *untimed student essays* are characterzied by a strikingly lower than average mean distance – in other words, the profiles form a denser cluster. The metadata that come with the respective ICE components hint at a stricter topical focus of these registers: the Jamaican *untimed essays* all deal with the topic "English for Academic Purposes: Tuition Fees" and the Canadian *timed essays* thematize English literature.

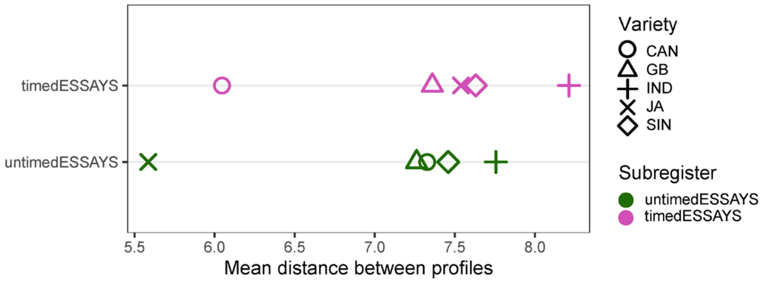


Figure 23: Average Euclidean distance between POS profiles of subregisters in *student writing*.

4.1.3. Visualization: Normalized frequencies

When encountering registers, subregisters or varieties that behave differently, as observed for the section *student writing* in Figure 22 and Figure 23, the intuitive question is which variables or tags are associated with the observed differences. In other words, how does the distribution of tags (be they MDA or POS) differ between the British, Singaporean, Canadian, Jamaican and Indian *student writing* sections. The visualization technique "Normalized frequencies" was written exactly for this purpose. It visualizes those tags that show the largest variation between the selected groups of data points. The grouping here again is determined by variety and register or subregister (depending on the selected level of granularity). Although ”Texts“ is an available option for **Level of granularity**, *ICEtree* uses “Subregister” internally for the identification of the most informative tags (see section 3.2.2 for details on how these are determined and which options are available). However, instead of plotting the normalized frequencies of the tags of the aggregated profiles, the normalized frequencies of individual profiles are plotted. This function is helpful in identifying outlier texts that might distort the aggregated profiles.

In parallel with other visualization techniques, the user is presented with some additional options in the submenu **View**, most of which are relatively self-explanatory and whose function is detailed in the documentation in the app itself.

To illustrate this visualization technique, the *student writing* sections of ICE-GB, ICE-CAN, ICE-SIN, ICE-JA and ICE-IND are selected. Figure 24 shows the top ten tags that differ most between the subregisters and varieties (**Order variables by** is set to “Mean decrease in Gini impurity (Random Forest)”). “Subregister” was chosen as level of granularity and is indicated by colour. Variety is indicated by different symbols. To exclude extremely infrequent tags, tags with a mean below 4 occurrences per 1,000 words were exlcuded using the functions in the submenu **Filter**. The X-axis in this plot indicates the normalized frequency per 1,000 words, the Y-axis the respective tags. The Canadian texts differ most prominently in the frequencies of the *s*-genitive (POS tag: GE) and the frequencies of 3rd person singular pronouns (POS tag: PPHS1). The Jamaican texts stand out through above-average frequencies of verbs in the infinitive (POS tags: VVI, VBI), modals (POS tag: VM), the *to*-infitive marker (POS tag: TO) and *for* tagged as preposition (POS tag: IF).

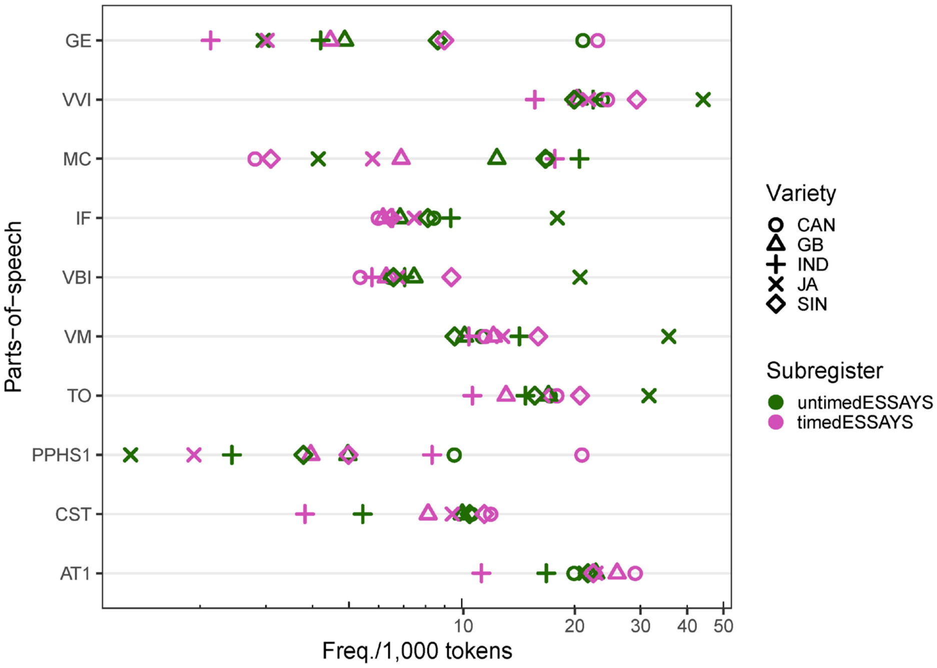


Figure 24: Normalized frequencies of tags that show the greatest difference between subregisters in *student writing* and varieties. Frequency threshold for variables was set to 4 per 1,000 words.

As the distribution of individual tags varies greatly, seeing patterns in the plot can become difficult in some cases. In those instances, the functions **Connect data points in dot plot** and **Show z-scores** can help. The first simply connects the data points with a thin line and the second tells *ICEtree* to standardize the observed tag frequencies. The X-axis then indicates z-scores. The interpretation of z-scores is as follows: a value of 0 indicates that the data point shows the mean frequency of the tag in question, a value of 1 or -1 indicates a tag frequency of one standard deviation above or below the mean respectively. Combined with *plotly*'s function to selectively hide data points, the tags in which the Canadian *timed essays* and Jamaican *untimed essays* differ from the other varieties are more clearly visible (Figure 25 right).

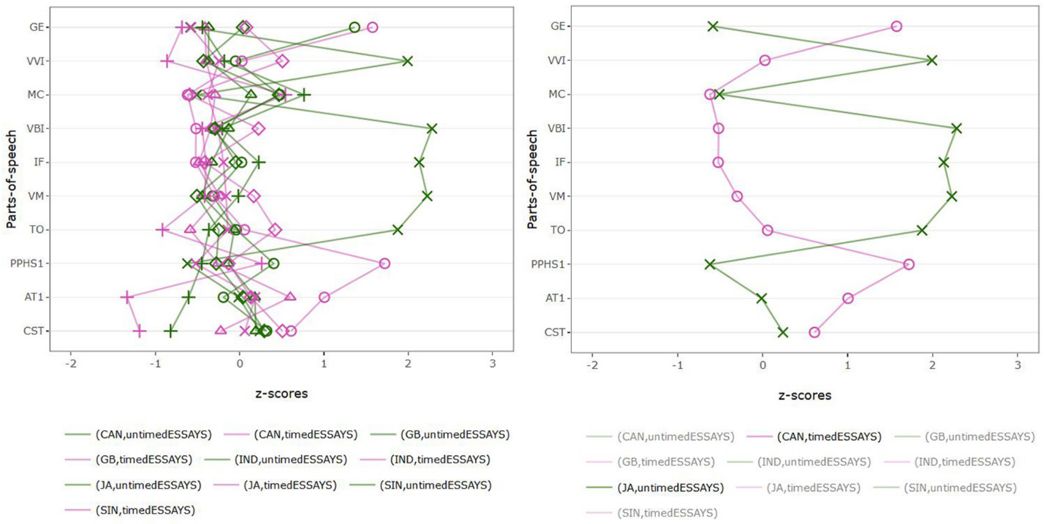


Figure 25: Z-scores of tags that show the greatest difference between subregisters and varieties. Frequency threshold for variables was set to 4 per 1,000 words. Left: Only data points for ICE-CAN *timed essays* and ICE-JA *untimed essays* are shown. Right: All data points visible.

4.2. A note on interpreting quantitative results

At this point, I wish to caution against a premature interpretation of the observed differences in frequencies of tags without cross-checking the quantitative results qualitatively. While the previously observed overuse of *s*-genitives in Canadian *student writing*, for example, might be connected to a topical focus on literature studies and the discussion of literary works and characters, a relatively high percentage of the genitive tags (GE) is indeed a result of faulty POS tagging. Single quotation marks that are separated by spaces from other words are tagged as genitive markers instead of punctuation marks. It is unclear, whether the student authors informing this section in ICE-CAN make heavy use of single quotation marks or whether this is an artefact of the transcription process. Utilizing concordance software, such as *AntConc*, and checking the metadata supplied with the corpora has proven to be invaluable when it comes to the interpretation and analysis of quantitative patterns observed in *ICEtree*.

4.3. Development and Dissemination

While specialized software, such as concordance programs, certainly have become indispensable tools in modern linguistic investigations, linguists also increasingly rely on a limited set of functions of programming languages for statistical analyses. This in itself is hardly a surprising development, given the speed of technological advancement and the apparently ever-growing size and complexity of databases available. To overcome the limitations of ready-made functions and software, however, a more profound knowledge of statistics and programming languages is necessary. Although the skill set of the future corpus linguist as portrayed in 2.4 comprises such abilities, it must be differentiated between using programming languages for one-time scenarios and writing programs for a wider community. The point is certainly not to downplay the former. Using programming languages for complex data manipulations and the development and application of new methods is key to the methodological advancement of the field. Examples of this category would be the works of Stefan Th. Gries and Douglas Biber. Where Gries (e.g. 2009, Gries & Mukherjee 2010) reports the results of experiments with a number of new corpus-linguistic methods, Biber (1988) used programming languages for the annotation of texts with a complex set of linguistic features and the subsequent factor analysis.

However, writing a program that is to be used by a wider community of users adds another layer of complexity entirely to the already intricate process of programming. For example, considerable thought must be given to the design of the user interface. It must be functional, intuitive and simple, but also include and allow the fine-tuning of more complex settings. Further, the programmer needs to anticipate errors and incorporate error handling. Errors might occur in the form of faulty input on the user-side, but may also be due to unexpected errors in the program itself. The first can be prevented by implementing checks of the user input. *ICEtree* does, for instance, verify that the user indeed selects corpus texts and tags and selectively shows and hides options which might lead to errors if combined. As errors due to faulty functions can never be fully ruled out, the programmer also needs to anticipate and handle those in the code as well and, if possible, provide the user with some feedback as to what happened. In those cases, the user might want to contact the developer of the program. Although this might sound trivial, in an academic context where employees often change their employer due to short-term contracts, the developer might want to set up a work-related email account not connected to a specific employer or university, or use platforms such as *GitHub* or the *Open Science Framework* (OSF)[28] for communication with their user base. Another issue is ensuring long-term availability and further development. Where the former can easily be achieved by relying on public platforms such as the OSF, the latter is necessary if bugs are identified in the program or if updates of other components that are required for the program to run are updated. In *R*, this is often the case when libraries on which the program depends receive major revisions. Additionally, a program requires a documentation where all the functions are described in detail. For *ICEtree*, this type of documentation can be found in the OSF repository and in the program itself. Another issue which needs to be addressed in this context is sadly also funding. If a program is developed in the context of a research project or a dissertation project such *ICEtree*, who funds and is responsible for the necessary continuous development and maintenance?

Connected to this is the last point I want to address in this section: the dissemination strategy of *ICEtree*. While I envisage an integration of the methods of *ICEtree* in existing corpus programs such as *CQPweb*, especially seeing that *ICE online* is in preparation at the University of Zürich, a standalone version of *ICEtree* and its source code will be released via the OSF. The link to the OSF

[28] The OSF is a public, free-of-charge platform that allows scientists to upload and share any kind of research related data, including manuscripts, presentations, data and source code.

repository, where a standalone version and the source code of *ICEtree* will be downloadable, is as follows: https://osf.io/ztfsx/.

Theoretically, the distribution strategies of shiny-based applications fall into one of two groups: either the app is made available via the browser (e.g. Wolk & Fastrich's 2019 *ShinyConc*), or the source code is published. Both paths again come with a few options. If the app is accessible via a web browser, a hosting solution is required. This can either be a commercial solution (e.g. https://www.shinyapps.io) or it can be self-hosted using RStudio's *Shiny Server* (https://rstudio.com/products/shiny/shiny-server/), which additionally requires the maintenance of a web server. If the app is distributed as source code, the end user is required to have an installation of *R* and all the required packages on their personal computer. Special attention must be given to the package versions as incompatibilities between various package versions are not uncommon in *R*, which in turn would result in *ICEtree* not working properly. Another possibility is to bundle the source code with a standalone version of *R* where all required packages are pre-installed in the right versions. Schlüter & Vetter (2020), for example, chose this publication strategy for their shiny-based application for the interactive exploration of the Google Books Ngrams data. *ICEtree* can also be downloaded as such a pre-bundled package via the OSF.

The downside of this publication strategy is the increased size of the bundle. Where the source code of *ICEtree* contains only a few megabytes, the download size of the bundled version amounts to more than 200 megabytes. This is due to the fact that the bundle includes a standalone version of *R* as well as required the packages. Yet, I consider this the best strategy for publishing shiny-based applications as it neither requires an internet connection for running the app, nor the installation of additional software, nor is the bundled version of the app susceptible to updates of the required *R* packages.

5. Investigating subregister variation with *ICEtree*

The following chapter utilizes *ICEtree* to investigate subregister variation in four registers of ICE. I use the term *subregister variation* here to emphasize that the focus lies on register variation within a corpus category on the lowest level of the sampling scheme, e.g. within *press editorials*. Although *ICEtree* is used for the case studies, the methodology applied in this chapter differs slightly from the one outlined in chapter 4, mainly due to the fact that manual classification of texts became necessary. Section 5.1 provides a detailed description of the methodology. Section 5.2. then proceeds to the analysis of the four registers *press editorials, skills and hobbies, business transactions* and *phone calls*.

5.1. Methodology

As I cautioned against a premature interpretation of quantitative results obtained from *ICEtree* in section 4.2, this chapter utilizes a mixed methods approach. In a first step, the quantitative methods available in *ICEtree* are used to explore whether texts of a single register in a given number of varieties form multiple clusters. In line with what is known about clustering texts by register based on a vector representation of the texts (see section 2.3), this has turned out to be a good indicator of texts of multiple subregisters being sampled into one register. To detect this, the visualization method is set to “Multidimensional Scaling” and the level of granularity to “Texts” in *ICEtree*. If texts cluster by variety as in Figure 26, the visualization “Normalized frequencies” can be applied directly.[29] However, if the patterns do not coincide with these pre-established groupings (Figure 27: left), manual reclassification becomes necessary.[30] Before the variables that differ between the clusters can be investigated, the registers of the respective clusters must be established – mainly through review of the metadata and facsimiles of the texts (where available) and close reading. If this step indeed reveals that texts in one register can be grouped into multiple subregisters (Figure 27: middle), the relevant situational characteristics are described and the texts are reclassified in the data selection tree in *ICEtree*. This way, the differences between the individual subregisters can be investigated further (Figure 27: right). These variables may then serve as a starting point for an in-depth functional analysis of linguistic features. However, as the overarching goal of this chapter is to illustrate how *ICEtree* can be used to detect instances where

[29] If no differences in the situational characteristics are detectable between the groups and other factors can be ruled out, the observed differences may very well point to regional disparities.

[30] For the user fluent in *R*, this is easily possible by modifying the main data table when running *ICEtree* from an IDE such as Rstudio. Simply navigate to the comment “# MANUAL TEXT CLASSIFICATION” in the file *config.R* and follow the instructions there. Adding such a functionality to the interface of *ICEtree* was beyond the scope of the first release candidate of *ICEtree*, but might be available in subsequent versions of *ICEtree*.

subregister variation might reduce comparability of components in ICE and not the exhaustive description of individual registers, this last step is omitted here.

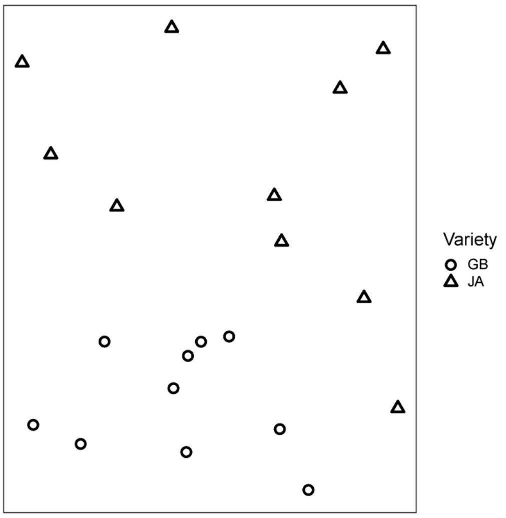


Figure 26: Texts of one register (*press editorials*) cluster by variety.

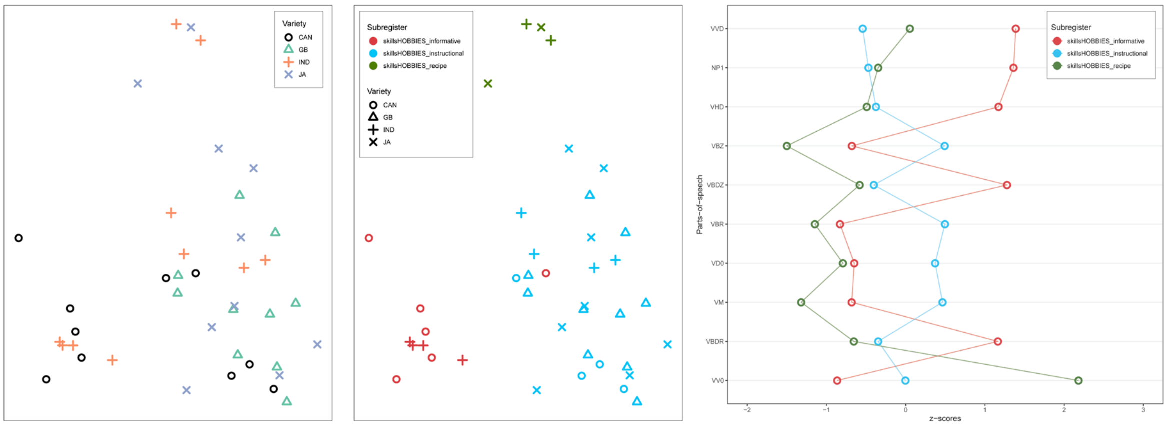


Figure 27: Illustration of ICEtree workflow with reclassification. Left: Texts in the subregister *skills and hobbies* form clusters. Centre: Texts reclassified by subregister. Right: Visualization "Normalized frequencies" for newly established text categories.

As the metadata of the press editorials alone were not sufficient to reliably determine the situational characteristics of the individual texts, manuals for and descriptions by practitioners of the field were also drawn on and facsimiles of the original texts of ICE-GB, ICE-CAN and ICE-JA were obtained where available. Furthermore, as it still could not plausibly be excluded that observed differences are indeed due to regional variation, additional corpus texts from the same newspapers and timeframes were sampled for the missing subregister for these three varieties (see 3.1.1.2 for details). In contrast to most other registers in ICE, vast archives of facsimiles are available for newspapers. Probably another consequence of the high historical and contemporary importance of this register and the large number of practitioners is that news writing is extensively studied and the characteristics are detailed in guides for prospective journalists. These resources are not available for less codified registers such as *skills and hobbies* or *business transactions*.

Although both POS and MDA profiles are available in *ICEtree*, only the POS profiles are investigated in this chapter as the focus lies on register variation and Vetter (to appear) shows that the MDA profiles are also sensitive to regional variation (see also p. 94).

5.2. Register variation between components of ICE

In this section, I will employ the methodology outlined above to explore individual registers in some components of ICE for register variation. As a brief revision, I will restate the underlying assumptions of the approach employed here:

i) Due to the non-operationalized definition of the registers included in ICE, different ICE teams sampled different subregisters into the same ICE register (see section 2.2).

ii) The opaque sampling process and the lack of metadata (again see section 2.2) render the identification of the individual subregisters or the reconstruction of the situational characteristics impossible.

iii) There is a strong connection between text types and registers so that register variation can be detected via clustering based on a large number of pervasive features (for detail see section 2.3), in this case the frequency of POS tags.

5.2.1. Press editorials

Even though not indicated in the ICE sampling scheme, the sections *press news reports* and *press editorials* could theoretically contain quite a number of different journalistic text categories, an overview of which is presented in Figure 28. In this context it is important to point out that these distinctions are not based on genre- or register-informed empirical investigations, but on the conceptualizations of practitioners and experts in the field. While the structure and number of categories may vary (e.g. Bell 1991: 12–17, Straßner 2000: 24–101, Cotter 2010: 143, Müller 2011: 323–385), opinion pieces are in general treated separately from news due to their lack of objectivity (cf. Cotter 2010: 60–61 on journalistic practice). In contrast to text categories in the news branch in Figure 28, which can occur in every section of the newspaper, texts in which the author foregrounds their personal opinion are traditionally confined to a dedicated *editorials* section or are explicitly labelled accordingly (Cotter 2010: 144–145).

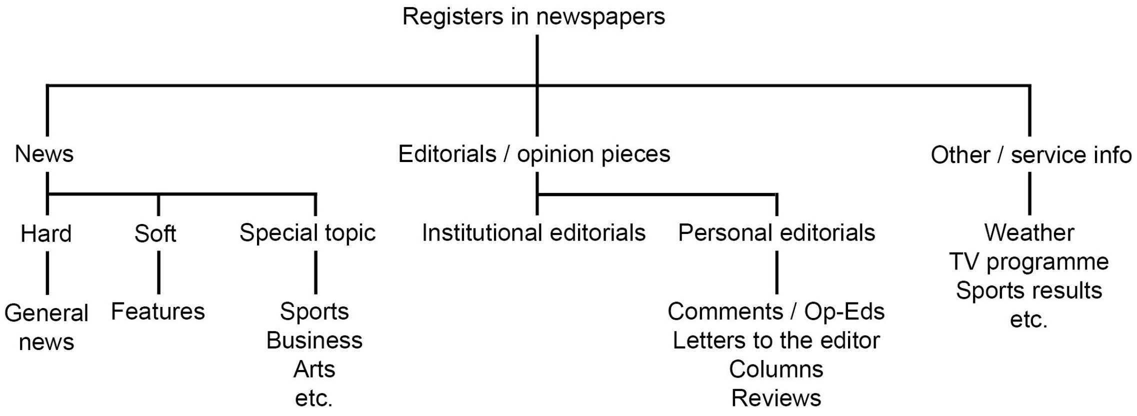


Figure 28: Registers in newspapers (adapted from Ljung 2000: 136).

Common text categories in the editorials section are institutional editorials (or *leaders/leading articles* in the UK), comments, letters to the editor (*L2E*), columns and reviews. The term *editorials* may be misleading as it can refer to either the section in newspapers where opinion pieces are printed, the opinion pieces which represent the opinion of the newspaper's editorial board (also institutional editorials) or all opinion pieces in a newspaper. In what follows, I will refrain from using *editorials*, and instead use the following terms for the sake of clarity: i) *institutional editorials* (*IE*) for editorials issued by and representing the opinion of the newspaper's editors, and ii) *personal opinion pieces* or *personal editorials (PE)* as cover terms for columns, comments and letters to the editor.[31] O*pinion pieces* serves as umbrella term for all opinionated newspaper text varieties (including institutional editorials).

IEs voice the opinion of the newspaper, are characterized by an institutional style and are written either by an editor-in-chief or a specialist for editorial-writing (Reeves & Keeble 2014: 35). These texts are sometimes printed beneath the logo or banner of the newspaper (Bell 1991: 13) and focus on current (often political) topics, but conventions may vary between local and national journalism, tabloids and quality newspapers (Wahl-Jorgensen 2009). *L2Es* often respond to opinion pieces published in a previous issue of a newspaper and play an important role in the marketing of the newspaper and the connection to its readership. It is therefore unsurprising that the choice of which letters are published also depends on the type of newspaper (local/national, quality vs. tabloid) (Richardson 2009). In comments, individual authors express their personal opinion on recent events or topics and adopt a first-person perspective accordingly. Authors of comments are not necessarily journalists or employed by the newspaper in which the comment is published. Columns take a special place in this section of newspapers as they are usually not tied to current events. Cotter (2010: 142) adds that columns further stand out as they permit a greater stylistic freedom and have a regular readership with which the author communicates more

[31] Reviews were excluded from this study as they can also occur outside the editorials section and are usually not referred to as *editorials*.

directly. McNair (2009: 115) writes that with the development of institutional editorials away from 'the proprietor's identifiable, direct mouthpiece' to their modern form, the collective, institutional voice of the newspaper, columns came to be the place where journalists could express their subjective opinions. Letters, columns and comments are usually signed and the latter two often include a picture of the author. There is, however, also some disagreement on the exact delimitations of some text varieties. Commonly, the boundary between comments and columns is drawn based on whether texts by the author are published regularly in the same space as is the case for columns (Reeves & Keeble 2014: 34–35). Other researchers (e.g. Müller 2011: 379) do not draw a clear distinction between these two text varieties.

5.2.1.1. Previous research on press editorials

While institutional editorials have gained the greatest attention of all opinion pieces, most linguistic studies focus on their argumentative and rhetorical structure (see Alonso Belmonte 2007 and Fartousi & Dumanig 2012 for an overview) and only few studies take a register perspective and investigate pervasive lexico-grammatical patterns. As this chapter adopts a comparative register perspective, the focus of the following overview will lie on the latter group of studies.

In comparison to other registers, Biber (1988: 127–164) finds that editorials[32] in the LOB corpus are characterized by a high type-token ratio, a high informational density, a relatively high frequency of modals, suasive verbs and infinitives but a relatively low frequency of past tense verbs and third person pronouns. Other studies (Sigley 2012, Morley & Murphy 2011) outline differences between institutional and personal editorials. Sigley (2012) finds systematic differences in the distribution of linguistic features associated with in-/formality between the two subregisters. He interprets his findings as personal editorials being less formal and abstract than their institutional counterpart. Morley & Murphy (2011: 209–210) find that institutional editorials exhibit higher frequencies of time adverbs, necessity modals, elements which help to structure the argument (e.g. *also, however, yet, rightly*) and expressions which 'indicate absolute certainty on the part of the writer and convey a sense of urgency' (e.g. *it is vital/clear that*). For personal editorials, the authors found high frequencies of first- and second-person pronouns, mental verbs (such as *think, know, think*) and adverbials of stance (e.g. *probably, certainly, actually* or *in fact*). Concerning institutional editorials, Westin (2002) reports a diachronic trends towards more informal, colloquial and less vague language for British institutional editorials – as opposed to Sigley (2012), who identifies the same trend only for American institutional editorials, but finds the opposite to be true for British editorials. Westin (2002) and Thompson (2014) further report stylistic differences between individual newspapers. Finally, cultural conventions also seem to have an effect on the realization of editorials (Bonyadi 2011).

[32] The editorials section in the LOB corpus contains *institutional editorials* as well as *personal editorials.*

5.2.1.2. Results

When including only the original ICE texts of ICE-GB, ICE-CAN and ICE-JA while ignoring the previously established subregisters in the *press editorials* section, the MDS solution in Figure 29 (left) shows a relatively clear tendency for the frequency profiles to cluster by variety. Although there is some overlap between the profiles of ICE-JA and ICE-CAN, the visualization suggests that there are systematic differences between press editorials in the varieties analysed. However, when classifying the texts into *PE*s and *IE*s (Figure 29: right), an altogether different pattern emerges. Except for two Jamaican outlier texts (JA:W2E-003, JA:W2E-010, indicated by broken circles), the POS profiles show a clear tendency to cluster by subregister.

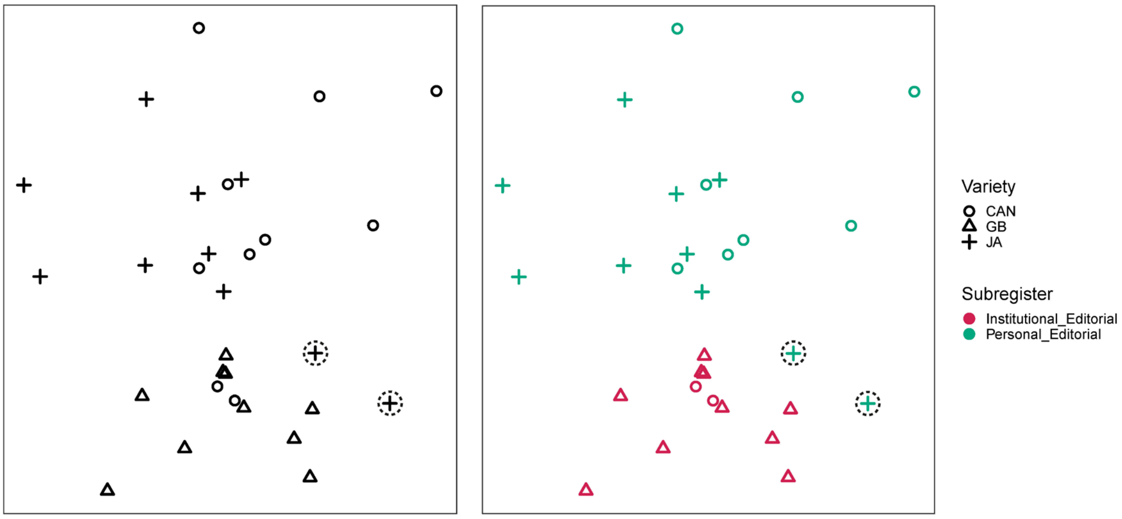


Figure 29: MDS plot of POS profiles of only the press editorials originally included in ICE-CAN, ICE-GB and ICE-JA (W2E-001-010). Left: No manual classification. Right: manual classification into *PE*s and *IE*s.

When including the newly sampled material (Figure 30), the pattern becomes even clearer and the tendency for the profiles to cluster by variety all but vanishes. A closer inspection of the two Jamaican outlier texts (indicated by broken circles) reveals possible reasons as to why they cluster with *institutional editorials* rather than the remaining *personal editorials*. The text file JA:W2E-003 contains two subtexts, the first of which includes long quotations from a document issued by a university administration not marked as extra-corpus material and therefore not excluded in the computation of the profiles. Although classified as *personal editorial*, the second subtext in this file is part of a series of highly argumentative texts that feature an impersonal style and discuss a constitutional reform. These texts are more akin to an *institutional editorial*. The text file JA:W2E-

010 contains two subtexts, one of which includes long quotations from a document issued by a government body, which again distorts the result.

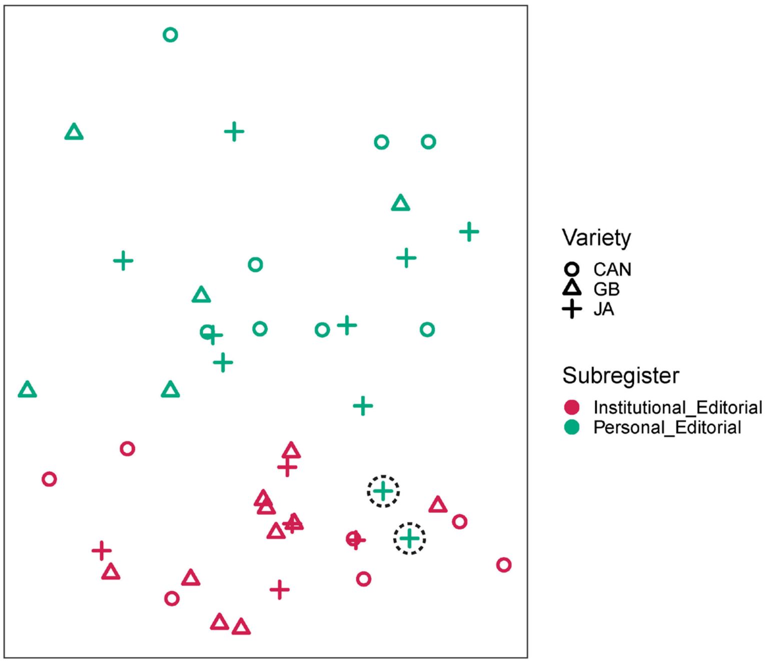


Figure 30: MDS plot of POS profiles of *press editorials* of ICE-CAN, ICE-GB and ICE- JA, including the newly sampled editorials (W2E-001-015).

Although the facsimiles were not retrieved for ICE-HK, ICE-IND and ICE-USA, close reading allowed me to also categorize the press editorials included in these varieties either into *IE*s or *PE*s. The result is shown in Figure 31. ICE-HK and ICE-IND contain exclusively and ICE-USA mostly *IE*s and the concrete realization of the texts (i.e. the text types) cluster with the varieties discussed before, further strengthening the overall picture.

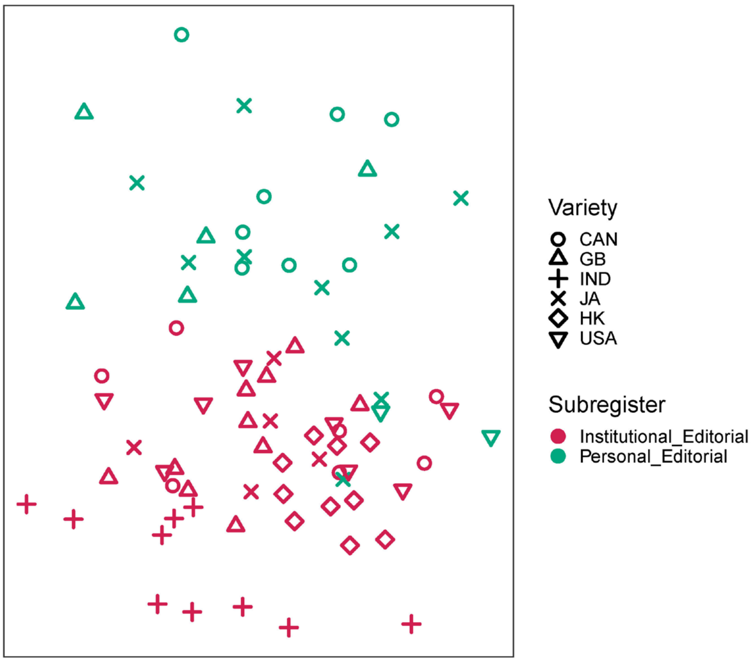


Figure 31: MDS of *press editorials* in ICE-CAN, ICE-GB, ICE-IND, ICE-JA, ICE-HK and ICE-USA. Texts were categorized into IEs and PEs.

A detailed analysis and discussion of the variables that differ between the British, Canadian and Jamaican *IE*s and *PE*s is presented in Vetter (to appear) – including a functional interpretation of the observed pervasive differences between two text types. In sum, the starkest contrast between the subregisters is drawn by the frequencies of pronouns (esp. *I*, *you* and *we*), contractions and modals. In line with a more subjective style, a high visibility of the author and a more direct connection to the readership, *PEs* exhibit high frequencies of *I* (PPIS1), *you* (PPY) *me* (PPIO1) and contractions. Vetter (to appear) argues that the observed features in *PE*s induce a colloquial tone and serve the effect of decreasing distance between the author and the reader. With respect to the frequencies of pronouns, this pattern holds when the three additional components ICE-USA, ICE-IND and ICE-HK are taken into account (Figure 32).[33] *IEs*, on the other hand, exhibit a larger number of modals and *be* as infinitive.

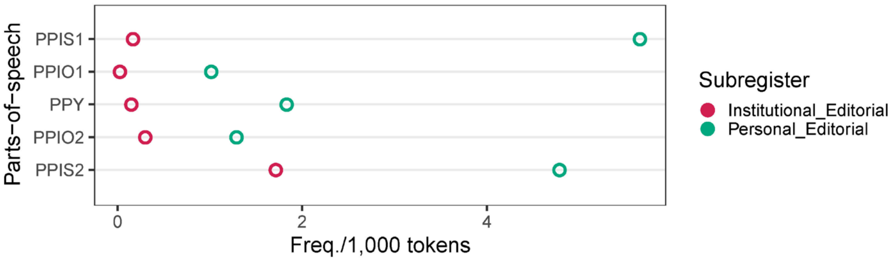


Figure 32: Top five variables which vary the most between *IE*s and *PE*s in ICE-HK, ICE-GB, ICE-CAN, ICE-USA, ICE-IND and ICE-JA. PPIS1 = *I*, PPIO1 = *me*, PPY= *you*, PPIS2 = *we*, PPIO2 = *us*.

[33] Modals and contractions were not investigated, as these are part of the MDA profiles.

5.2.2. Skills and hobbies

The analysis of the ICE register *skills and hobbies* was triggered by an internal discussion in the context of the compilation of ICE-PR and ICE-MT. Since finding material for some registers in L2 varieties can be extremely challenging, the central point of the discussion was whether texts in *skills and hobbies* must indeed offer instructions – *skills and hobbies* is a subregister of *instructional* writing after all – or whether purely informative texts about hobbies (e.g. a travel blog, if the author considers travelling a hobby of theirs) are also acceptable. Although Nelson (1996: 33) writes that texts in this section "offer instruction [...] [and] include publications such as car manuals, cookery books, and gardening manuals", spot checks of texts in the form of close reading revealed that not all texts in some ICE varieties offer instructions. Some are purely informative: In the three subtexts in the corpus file ICE-CAN:W2D-011, for example, the authors write about their experience of watching reindeers in the arctic (ICE-CAN:W2D-011:1), describe the atmosphere during the Arctic Winter Games in the Northwest Territories (ICE-CAN:W2D-011:2) and portray a certain athlete (ICE-CAN:W2D-011:3). The subtexts in ICE-IND:W2D:013 are an homage to a late Indian musician and the description of a concert held in honour of his death (ICE-IND:W2D-013:1) and three texts praising the performance of musicians during local concerts (ICE-IND:W2D:013:2-4). While the topics in the instructional texts in *skills and hobbies* are equally diverse – ranging from tips on keeping one's dog healthy (ICE-CAN:W2D-012), to buying and maintaining houses (ICE-GB:W2D-012) or changing the drive belt of a car (ICE-GB:W2D-018) – these two groups of texts differ in one situational characteristic: communicative intent.

This section illustrates the use of *ICEtree* that ultimately led to the reclassification of texts in *skills and hobbies* in ICE-CAN, ICE-JA, ICE-GB and ICE-IND into instructional texts *(skillsHOBBIES_instructional)*, informative texts *(skillsHOBBIES_informative)* and cooking recipes *(skillsHOBBIES_recipe)* (see also section 3.1.1.2).

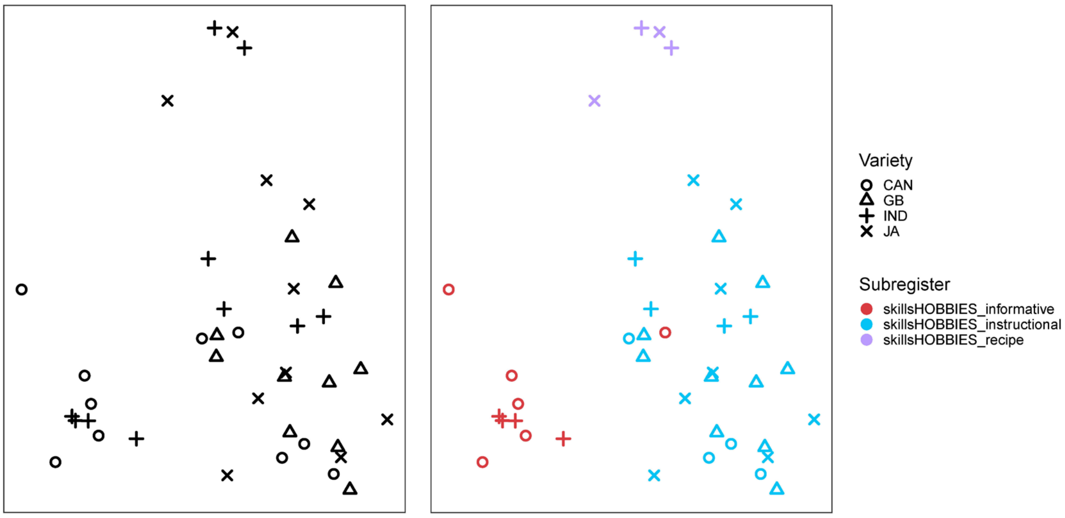


Figure 33: Comparison of MDS of texts in category *skills and hobbies* in ICE-CAN, ICE-GB, ICE-IND and ICE-JA with and without subregister classification.

Both panels in Figure 33 show the MDS plot based on the POS profiles of all texts in skills and hobbies in ICE-CAN, ICE-JA, ICE-GB and ICE-IND. In this MDS solution, three distinct clusters are visible: a large cluster located at the bottom right, a smaller one comprising Canadian and Indian texts at the bottom left and a small one comprising two Jamaican and two Indian text at the top of the plot. In order to investigate whether the variation in text types coincides with subregister variation, the subregisters are indicated by colour (right panel). With the exception of one Canadian text (W2D-016, two reviews of snowmobiles), the text types overlap clearly with the register classification. The difference in situational characteristics (i.e. communicative purpose) is clearly reflected in the distribution of pervasive features. An influence of variety on the MDS solution is not visible.

As a result of the reclassification of the texts, *ICEtree*'s function "Normalized frequencies" can be used to explore which POS tags vary the most between groups. The advanced settings were used to increase the number of variables to 15 as the top six variables almost exclusively consisted of POS tags relating to verbs. Figure 34 shows the z-scores of the top 15 features with the greatest variation between the registers.

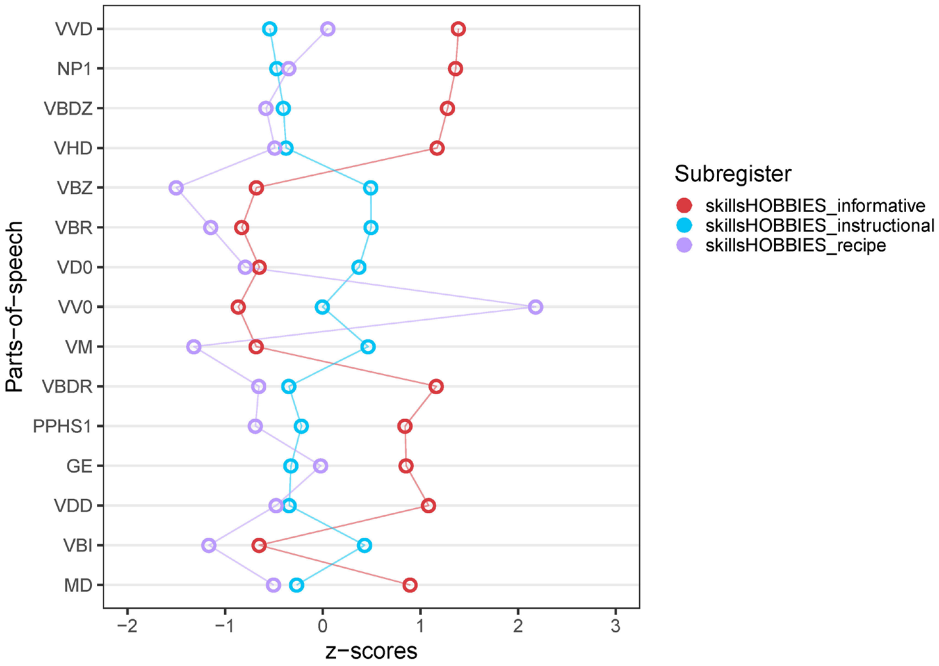


Figure 34: Top 15 POS tags that show the greatest differences between the subregisters, *ICEtree* function "Merge varieties (i.e. conflate corpora)" enabled.

Compared to the instructional texts (*skillsHOBBIES_instructional*), the informative texts (*skillsHOBBIES_informative*) exhibit significantly higher frequencies of past tense verbs (POS tags VVD, VBDZ, VHD, VDD), 3rd person singular pronouns (PPHS1), singular proper nouns (NP1, mostly names and place names) and ordinal numbers (MD, mostly used in a sports context). These features typically co-occur when an author writes about past events or tells a story. In the text CAN:W2D-017, for example, the author writes about a fishing trip he went on. While fishing appears to be a personal hobby, the text contains no instructional elements. Instead, the communicative purpose is likely to entertain the readership with a detailed recollection of the events of this experience.[34] The following excerpt of this text illustrates quite nicely how the features are used for this purpose:

(1) Thirty years from now, when I have a grandchild on each knee, both eager to hear tall tales, I'll probably tell them about the three days I spent on Quebec's Ste. Marguerite River in 1989. My tale will begin innocuously enough. I'll tell them that I had just flown back from an unsuccessful trip to the Kegaska, a pretty-as-a-picture river on Quebec's lower North Shore, and that I was a little tired when I climbed into my car, at six in the morning, for the eight-hour drive from Montreal to Sacre-Coeur. But the sun was out, it was warm, and I was off to fish a river I'd never seen before. That alone would have been enough to lift my spirits nine times out of 10. On this trip, however, my angling companions were to include Lee and Joan Wulff, both fly-fishing legends in their own time. (ICE-CAN:W2D-017#2-7:1, subregister: skillsHOBBIES_informative)

Instructional texts (*skillsHOBBIES_instructional*) excluding the recipes are characterized by higher frequencies of present tense verbs (VBR, VBZ, VD0, VV0), modals (VM) and the verb *be* in its infinitive (VBI). The following examples serve to illustrate the use of these features and are taken from a text that offers advice for dog owners on taking care of their dog's health (ICE-CAN:W2D-12) and a text offering guidance for potential house buyers. Present tense verbs are often used either in imperatives as in (2) or in factual statements as in (3) or (4).

(2) As you investigate your choices, **find out** if the breeds you are interested in are pre-disposed to any health problems such as hip dysplasia or congenital heart disease. (ICE-CAN:W2D-012#13:1, subregister: skillsHOBBIES_instructional)
(3) Attention to good health **begins** as you decide what breed of dog best fits your needs. (ICE-CAN:W2D-012#11a:1, subregister: skillsHOBBIES_instructional)
(4) The hot water cylinder **supplies** the domestic hot water to the taps. (ICE-GB:W2D-012#98:1)

Modals are often used to indicate what the reader *must* or *should* do (e.g. (5) and (6)), in questions to encourage the reader to reflect on something (e.g. (7) and (8)), to indicate what *will* be the outcome of a certain kind of action as in (9) or to indicate the ideal state of affairs as in (10).

(5) The next visit **should** be for neutering. (ICE-CAN W2D-012 #71:1, subregister: skillsHOBBIES_instructional)

[34] Although the author might use this story for didactic purposes later in the text, only the first 2,000 words are included in the corpus, and this part of the text does not exhibit an instructional character. This is connected to the issue that different parts of texts may serve different communicative purposes.

(6) Although an occasional treat from the table isn't likely to harm your dog, you **must** be careful. (ICE-CAN W2D-012 #86:1, subregister: skillsHOBBIES_instructional)

(7) **Can** you afford the running costs and maintenance costs? (ICE-GB:W2D-012#58:1, subregister: skillsHOBBIES_instructional)

(8) **Would** human food do just as well? (ICE-CAN W2D-012 #84:1, subregister: skillsHOBBIES_instructional)

(9) He or she **will be** the second most important person in your dog's life. (ICE-CAN W2D-012 #17:1, subregister: skillsHOBBIES_instructional)

(10) And remember, flues should not be interconnected with any other room. (ICE-GB:W2D-012#70:1, subregister: skillsHOBBIES_instructional)

Compared to *skillsHOBBIES_instructional*, *skillsHOBBIES_recipe* are further characterized by relatively high frequencies of present tense lexical verbs (VV0), excluding 3rd person singular forms (e.g. "he/she/it thinks", the corresponding POS tag would be VVZ). Additionally, representative excerpts for each subregister are provided at the end of this section.

Spot checks of the POS tags with the concordance software *AntConc* (Anthony 2018) revealed another interesting pattern: more than half of the lexical verbs tagged as VV0 in *skillsHOBBIES_recipe* and approximately a third in *skillsHOBBIES_instructional* occur in sentence-initial position (see Figure 35). Sentence-initial position was determined based on whether the first letter of the verb is in capital letters. The regular expressions used in *AntConc* were thus "[A-Z]\w+_VV0" for sentence-initial VV0s and "[a-z]\w+_VV0" for verbs occurring within the sentence. Except for some mistagged words, spot checks confirmed the overall accuracy of these search expressions. Sentence-initial VV0s are often part of short and consecutive imperatives that instruct the reader in a step-by-step manner, e.g.:

(11) Add enough warm water to mix a soft dough. (ICE-IND:W2D-015#74:1, subregister: skillsHOBBIES_recipe)

(12) Slacken the front wheel bolts, raise and SAFELY support the front of the car and remove the front wheels. (ICE-GB:W2D-018#37:1, subregister: skillsHOBBIES_instructional)

(13) Start by asking your friends and neighbours who have pets where they go, whether they are satisfied, and why. (<ICE-CAN:W2D-012#44:1, subregister: skillsHOBBIES _instructional)

In contrast, this highly pervasive register marker is almost entirely absent in *skillsHOBBIES_informative*. The frequency of verbs tagged VV0 is dramatically lower in informative texts and the vast majority does not fill a sentence-initial position.

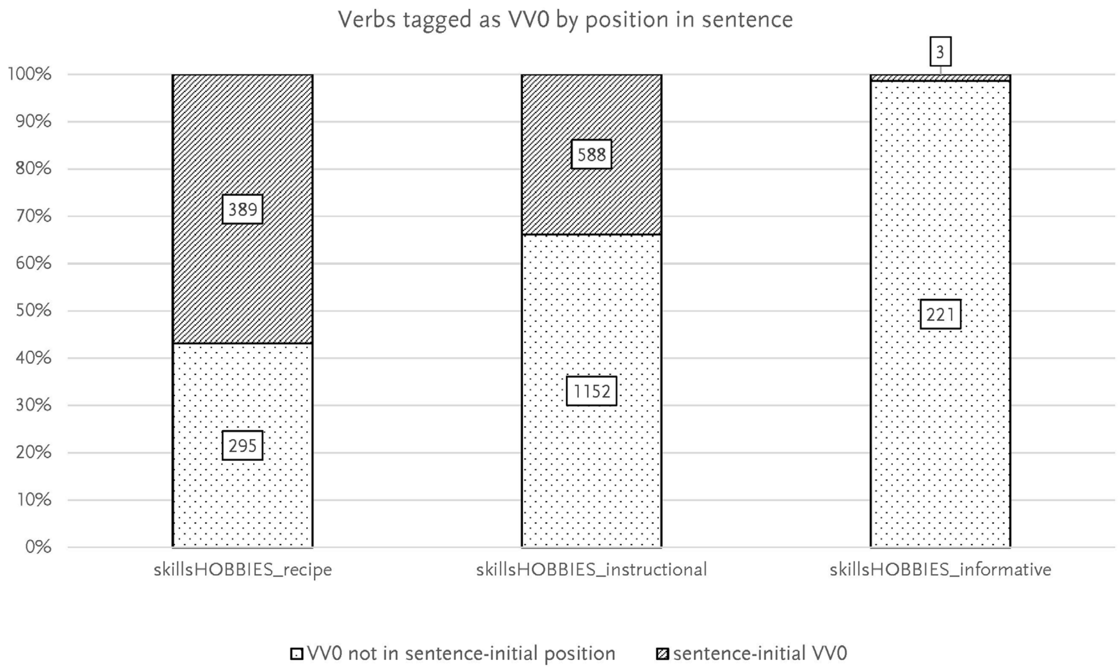


Figure 35: Sentence position of verbs tagged VV0 in percent. Labels indicate the absolute number of occurrences.

Excerpt of an informative text in skills and hobbies

<ICE-CAN W2D-011#70:2> Sport In a Cold Climate BY JUDY LANGFORD
<ICE-CAN W2D-011#70a:2> Two things set the Arctic Winter Games apart from most other sports competitions- weather and attitude.
<ICE-CAN W2D-011#71:2> During the NWT playdowns for this year's Games, the temperature hit minus 45&degree; Celsius.
<ICE-CAN W2D-011#72:2> 'It's so cold, the gunpowder doesn't burn properly,' groaned a silhouette shooter just in from the deepfreeze.
<ICE-CAN W2D-011#73:2> 'You're never sure where your shots are going to go.'
<ICE-CAN W2D-011#74:2> As he huddled over a woodstove, another competitor asked, 'Is your hat on fire?'
<ICE-CAN W2D-011#75:2> 'No.'
<ICE-CAN W2D-011#76:2> It's steaming because I came inside.'
<ICE-CAN W2D-011#77:2> All the shooters' hats, scarves and beards were caked with ice.
<ICE-CAN W2D-011#78:2> Their hands were so stiff they could barely reload their weapons.
<ICE-CAN W2D-011#79:2> In fact, the temperature was considered too cold for Snowshoe Biathlon.
<ICE-CAN W2D-011#80:2> But the rulebook has provisions for such extremes.
<ICE-CAN W2D-011#81:2> At Hay River, in a sort of virtual biathlon, the athletes ran in place in a gymnasium, then ducked outside to shoot targets.

<ICE-CAN W2D-011#82:2> The dog mushers delayed their competition by one day, partly due to the cold and because bad weather had delayed a team flying in from the Eastern Arctic.
<ICE-CAN W2D-011#83:2> Officials said they didn't want anyone to miss their chance.

Excerpt of a recipe in skills and hobbies

<ICE-JA:W2D-012#1:1> <h> Alligator Pear Relish </h>
<ICE-JA:W2D-012#2:1> Lady Nugent, wife of General George Nugent, Governor of Jamaica from 1802-5, refers to avocado as <quote> 'subaltern's butter' </quote> in her Journal.
<ICE-JA:W2D-012#3:1> The impecunious junior officers of the British Army, stationed in Jamaica, would use this cheap fruit to spread on their bread instead of using butter, which was expensive and difficult to keep in the tropical Jamaican climate prior to refrigeration.
<ICE-JA:W2D-012#4:1> 1 Ripe avocado 1
<ICE-JA:W2D-012#5:1> 1 tbsp Lime or lemon juice 15 ml
<ICE-JA:W2D-012#6:1> 1/4 tsp Salt 1 ml
<ICE-JA:W2D-012#7:1> 1/8 tsp Freshly ground black pepper 5 ml
<ICE-JA:W2D-012#8:1> 1 tbsp Busha Browne's Original Spicy Planters Sauce 15 ml
<ICE-JA:W2D-012#9:1> Dash Busha Browne's Pukka Hot Pepper Sauce Dash
<ICE-JA:W2D-012#10:1> Crush avocado lightly and immediately sprinkle with lime or lemon juice to prevent a change of colour.
<ICE-JA:W2D-012#11:1> Add the salt, pepper, Planters Sauce and Pukka Sauce.
<ICE-JA:W2D-012#12:1> Mix lightly with fork.
<ICE-JA:W2D-012#13:1> Serve at once with breadfruit chips, pita bread triangles or crispcrackers or toast.

Excerpt of an instructional text in skills and hobbies

<ICE-GB:W2D-018#1:1> <h> <bold> DRIVE BELTS </h> </bold> &dotted-line; 12
<ICE-GB:W2D-018#2:1> On the OHV engine both the alternator and water pump are driven by a &lsquo; V &rsquo; belt, but on the OHC engine only the alternator is driven by an external belt from the crankshaft pulley, in this case a &lsquo; Poly-Vee &rsquo; multi-groove belt, although models with power steering also use this belt to drive the steering pump.
<ICE-GB:W2D-018#3:1> It is important that the correct tension of the &lsquo; V &rsquo; or &lsquo; Poly-Vee &rsquo; drive belt is maintained to ensure efficient operation of the electrical (and in the case of the OHV engine, cooling) system.
<ICE-GB:W2D-018#4:1> Too great a tension will place excessive strain upon the alternator or water pump bearings and cause undue wear on the belt.

<ICE-GB:W2D-018#5:1> To test the belt tension, press the belt down at a point midway on the longest run between pulleys (Fig. A:25), using firm thumb pressure.
<ICE-GB:W2D-018#6:1> The belt should deflect 13 mm (0.5 in).
<ICE-GB:W2D-018#7:1> If retensioning of the belt is necessary, you will first have to remove the inner wing access plate on OHV models by undoing the two retaining screws, unclipping the plate at its leading edge and pulling it downwards (
<ICE-GB:W2D-018#8:1> Fig. A:26).
<ICE-GB:W2D-018#9:1> The plate is held at the bottom by a plastic pin which need not be undone as the plate can be flexed out of the way.
<ICE-GB:W2D-018#10:1> On both engine types, slacken the alternator mounting bolts / (Figs. A:27 &ampersand; A:28).
<ICE-GB:W2D-018#11:1> Pull or lever the alternator away from the engine, applying any force to the drive end bracket only, until the correct tension is obtained.
<ICE-GB:W2D-018#12:1> Note that some OHC engines have a tension adjuster screw (A, Fig.28) and this can be used instead of levering the alternator.
<ICE-GB:W2D-018#13:1> Tighten the alternator adjuster slide bolt before tightening the mounting bolts.
<ICE-GB:W2D-018#14:1> Recheck the tension and readjust as necessary.
<ICE-GB:W2D-018#15:1> Refit the access plate on OHV models.

5.2.3. Business transactions

As described in section 3.1.1.2, texts in the category *business transactions* in ICE-GB, ICE-CAN, ICE-IND and ICE-JA were reclassified into three distinct subregisters based on a difference in situational characteristics. The first subregister consists of interactive discussions in a business context (*businessTRANS discussion*). Texts in this subregister include planned and more formal discussions in a business context where participants represent institutions (*businessTRANS institutional*) and reports brought forth in business meetings (*businessTRANS report*). The interactive discussions include, for example, a sales conversation where a customer wants to buy a pair of skis and seeks professional consultation (ICE-CAN:S1B-074), a discussion between an architect and two clients (ICE-GB:S1B-071), a faculty meeting at the University of London (ICE-GB:S1B-075) and a meeting of an Indian educational board (ICE-IND:S1B-071).

Three distinct clusters are visible in the MDS solution of all texts in business transactions in ICE-GB, ICE-CAN, ICE-JA and ICE-IND (Figure 36). The first cluster is comprised of all Canadian, British, two Indian (ICE-IND:S1B-071 & ICE-IND:S1B-075) and two Jamaican (ICE-JA:S1B-071 & ICE-JA:S1B-072) texts, the second of the remaining Jamaican texts and the third contains the remaining Indian *business transactions* (although two texts are positioned between both clusters).

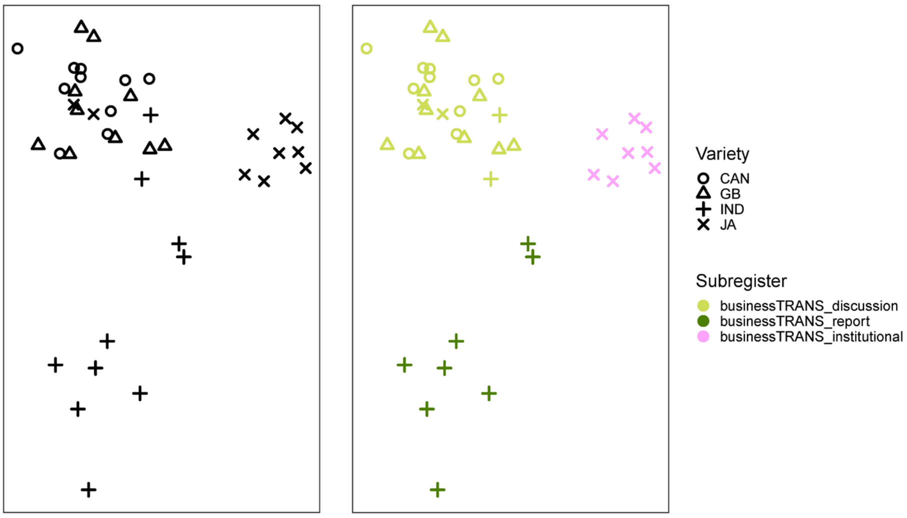


Figure 36: MDS of texts in category *business transactions* in ICE-CAN, ICE-GB, ICE-IND & ICE-JA with and without subregister classification.

Although most texts in ICE-JA form a distinct and dense cluster, the metadata accompanying the corpus indicate that the sources of the institutional texts are verbatim notes suggesting discrepancies in the transcription practices for these texts. Close reading showed that the mark-up differs from other spoken texts in the corpus as there are no indications of overlapping speech, pauses and hesitation markers. As hesitation markers are part of the POS profiles, the relevant POS tag is excluded in the further analysis. Figure 37 shows a new MDS solution with hesitation markers and interjections (POS tag UH) excluded from the frequency profiles. While the Jamaican texts still manifest as a denser cluster (probably due to the topical focus), only two large clusters remain: one containing all texts of ICE-JA, ICE-CAN, ICE-GB and the two previously mentioned texts of ICE-IND. Except for the two Indian outlier texts that are positioned in between, the other cluster contains the remaining texts of ICE-IND.

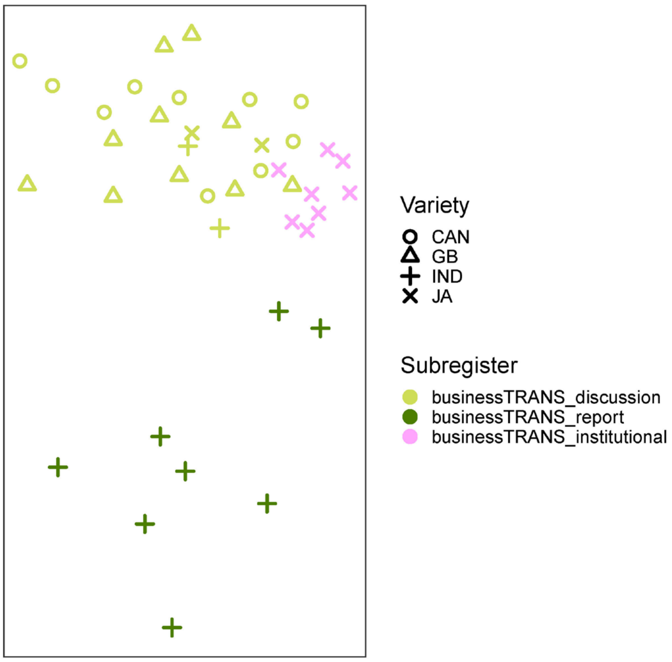


Figure 37: MDS of texts in category *business transactions* in ICE-CAN, ICE-GB, ICE-IND & ICE-JA. Subregisters indicated by colour. POS tag UH excluded.

Although the interactive and institutional business discussion differ text-externally and also text-internally (in terms of minimal responses, the use of pronouns and average length of utterances), this is not visible in the second POS-based MDS solution. However, it is unclear whether the lack of minimal response and the greater length of (uninterrupted) utterances in institutional texts is a result of the transcription practices. While beyond the scope of this study, the sampling of additional material of these subregisters could shed light on whether there is indeed a difference in text types detectable with POS profiles. The POS tags that vary the most between the subregisters are shown in Figure 38. For most POS tags, the interactive discussions and the reports behave complementarily. Where the latter is characterized by a high frequency of articles (AT), general prepositions (II), singular common nouns (NN1), past participles of lexical verbs (VVN), the discussions exhibit higher frequencies of adverbs (RR), first person singular subjective pronouns (PPIS1), lexical verbs in the infinitive (VVI), *do* (VD0), *that* tagged as a conjunction (CST) and negations (XX).

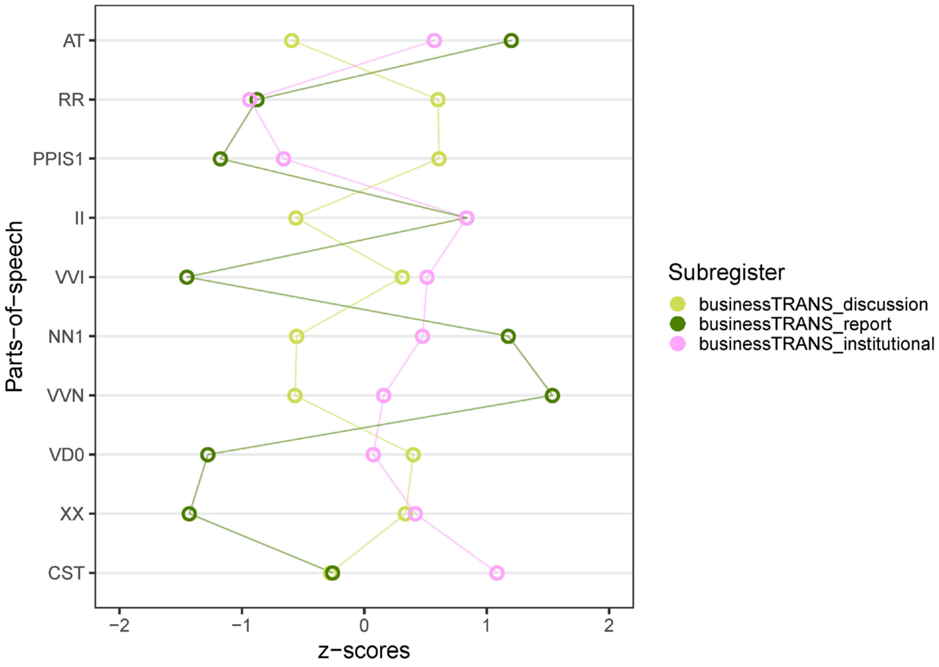


Figure 38: Top 10 POS tags that show the greatest differences between the subregisters. ICEtree function "Merge varieties (i.e. conflate corpora)" enabled.

High frequencies of articles, nouns, prepositions and past tense verbs fit well with the situational characteristics of the reports in business meetings, where the focus is on presenting information of past events (e.g. minutes of a previous meeting) in an impersonal style and with an explicit frame of reference. Hence, noun phrases are often realized as simple noun phrases consisting of proper nouns or titles (e.g. (14)) or as complex noun phrases including postmodifying prepositional phrases (e.g. (16)), adverbials are realized as prepositional phrases detailing precise information of time and place (e.g. (16)) and sentences often stand in the passive:

(14) The members were also informed that Vice-chairman and honorary treasurer were not in town (ICE-IND:S1B-073#18:1:A, subregister: businessTRANS_report)

(15) Then uh the regular agenda was taken up for discussion (ICE-IND:S1B-073#19:1:A, subregister: businessTRANS_reports)

(16) One confirmation of minutes the minutes of the meeting held on twenty-fifth December nineteen ninety-three were read and confirmed (ICE-IND:S1B-073#20:1:A, subregister: businessTRANS_institutional)

The features characteristic of the discussions highlight their similarity to spoken conversations. These features are used by interlocutors to express their personal views and engage in lively discussions (e.g. via the use of *do*-support for emphasis in (19) or subjuncts as in (17)) with a more situation-dependant frame of reference as is evident from the frequent use of pro-forms (e.g. in (18)).

(17) You know there's the Brown Corpus but I don't think that Brown provides that actually (ICE-GB:S1B-076#93:1:A, subregister: businessTRANS_discussion)

(18) No I don't think it does either (ICE-GB:S1B-076#94:1:B, subregister: businessTRANS_discussion)

(19) And the only information source I know that does provide it was done in the thirties manually by a chap called Irving Lorge (ICE-GB:S1B-076#95:1:A, subregister: businessTRANS_discussion)

The institutional business transactions take a middle ground, which is also evident from the texts (see below). While still discussions in a business context, these texts are less interactive, have a strong focus on contents and are more planned than texts in the subregister *businessTRANS discussion*. Representative text excerpts are provided at the end of this section.

Excerpt of institutional business transactions

<ICE-JA:S1B-078#25:1:A> We have an interest in resolving the loan situation and of course we have an interest in the University of not subsidizing the bank
<ICE-JA:S1B-078#26:1:B> Surely and it may require the coming together of the main players working out the plan for that individual going to the bank X bank Y bank working out an individual arrangement to determine the stronger position as against the other situations in which you go with a block of business and you would be more able to negotiate a stronger point
<ICE-JA:S1B-078#27:1:B> It also gives the University some sense of what is the exposure atany point in time
<ICE-JA:S1B-078#28:1:A> It has to be settled across the table it cannot leave this negotiation
<ICE-JA:S1B-078#29:1:A> Our experience is that we are back here in three years without an agreement
<ICE-JA:S1B-078#30:1:A> We are flexible we would want to see your proposal on the table see plan B between now and tomorrow
<ICE-JA:S1B-078#31:1:A> It must be a smaller group outside at the end of the day when we have to sign off on those things

Excerpt of report in business transactions

<ICE-IND:S1B-077#1:1:A> So <,,> the last meeting <,,> which perhaps was the first <,,> special meeting <,,> requisition <,,> was held on nineteenth April nineteen ninety-four <,,> at three pm in the physics auditorium <,,> to discuss <,,> non issuance of <,,> convocation invitations to non <O> one word </O> teachers <,,> of this <,> University <,,>
<ICE-IND:S1B-077#2:1:A> In the meeting <,,> after <,,> the relevant documents were presented <,,> and <,> after much of discussion about the issue involved <,,> it was

resolved <,,> that in future <,,> the attempt <,> should be made by the organisation <,,> to convene the authority <,,> that the teachers of the University <,,> shall be issued <,,> invitations for the convocation <,,> that is all convocations to be <,> held in future <,,>
<ICE-IND:S1B-077#3:1:A> Now <,> that was the only item to be discussed in that meeting <,> and that constitutes the minute <,> of the previous meeting <,,>
<ICE-IND:S1B-077#4:1:A> The meeting <,,> preceeding this special meeting <,,> was held on fifteenth January nineteen ninety-four <,,>
<ICE-IND:S1B-077#5:1:A> In that meeting <,,> apart from the confirmation to the previous <,> meeting <,,> minute <,,> in all <,,> twenty resolutions were passed <,,>

Excerpt of discussion in business transactions

<ICE-CAN:S1B-077#1:1:A> Okay
<ICE-CAN:S1B-077#2:1:A> So on the fourteenth we have a staff meeting <,>
<ICE-CAN:S1B-077#3:1:B> Yup <,,>
<ICE-CAN:S1B-077#4:1:A> Last week's meeting <,,>
<ICE-CAN:S1B-077#5:1:A> <}> <-> I'll tell </-> <=> I'll fill </=> </}> you in a bit
<ICE-CAN:S1B-077#6:1:A> We started <}> <-> <.> talk </.> </-> <+> talking </+> </}>
<ICE-CAN:S1B-077#7:1:A> It was a <unclear> maybe two words </unclear> meeting right <{> <[> <,,> </[> as opposed to a team staff meeting
<ICE-CAN:S1B-077#8:1:B> <[> Ya </[> </{>
<ICE-CAN:S1B-077#9:1:A> So uhm <,> <@> Gerald </@> uh <,> chaired and we were talking about protocol and <indig> fonction d'&eacute;quipe </indig> and stuff
<ICE-CAN:S1B-077#10:1:A> Anyway he was just <}> <-> saying </-> <=> commending </=> </}> us he says that <}> <-> it's </-> <=> <,> since </=> </}> our last meeting it's been going very well <,> you know between the two offices
<ICE-CAN:S1B-077#11:1:B> That's nice he had something <,> positive to say
<ICE-CAN:S1B-077#12:1:A> Ya
<ICE-CAN:S1B-077#13:1:A> And then uhm <,> but he said you know there's still a lack of communication between you know <@> Tori </@> and I <,> so that we have to work on
<ICE-CAN:S1B-077#14:1:A> So we talked about like any problems and stuff
<ICE-CAN:S1B-077#15:1:A> And <,> <@> Amelie </@> brought up the fact that <,> when we sat down <,> that Friday <,,> to talk about the <{> <[> <,> <?> dossier </?> </[> <}> <-> how good </-> <=> how interesting </=> </}> it was

5.2.4. Phone Calls

Depending on the relationship between the speakers, I divided the texts in the category *phone calls* into *private* and *public phone calls* (for detail see section 3.1.1.2). It is noteworthy that ICE-JA contains only *public phone calls*, whereas the other varieties contain only *private phone calls*. When subjecting the distances between POS profiles to MDS (Figure 39), the two subregisters in phone calls form two clear-cut clusters. The first comprises all *private phone calls*, the second the *public phone calls*.

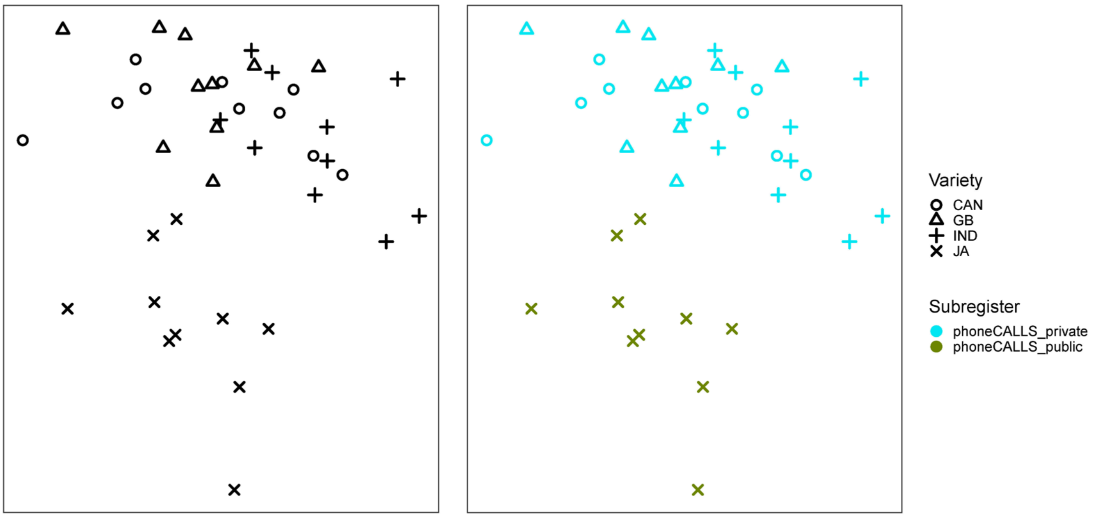


Figure 39: MDS of texts in category *phone calls* in ICE-CAN, ICE-GB, ICE-IND & ICE-JA with and without subregister classification.

Although the previous sections illustrated that regional variation appears to have little effect on the distances between POS profiles, it cannot be entirely ruled out in this case as the ICE components are homogenous in their composition in terms of sampled subregisters. Nevertheless, the POS which differ most between the two subregisters are shown in Figure 40. The most drastic differences can be observed in the frequencies of articles (AT), singular common nouns (NN1) and interjections (UH).[35] Possible reasons for these differences might be that phone calls to hotlines or radio shows have a stronger focus on conveying contents and are less spontaneous and more planned than private phone calls. This is also mirrored in the higher frequency of articles, nouns and lower frequency of 3rd person singular nouns (PPHS1), which point to a more explicit style and that shared knowledge is negotiated (as is evident in the excerpt of the hotline call at the end of this section). Additionally, the social distance between speakers is higher as they are often strangers[36] and they also speak for the benefit of an unenumerated

[35] Concerning interjections, discrepancies in transcription practices as cause seem unlikely as the transcriptions of the Jamaican *public phone calls* include a wide variety of mark-up, including overlapping speech, interjections, minimal responses and fillers.

[36] Upon reading the transcripts of *public phone calls*, it becomes evident that the caller and the operator or radio show host share a very one-sided relationship akin to that of column writers and their readers. The listeners or callers appear to feel a personal relationship with the host, yet this relationship is characterized by a lack of shared personal knowledge and contact. This cannot be compared with a relationship between family, friends or even colleagues – hence I prefer to use the term *stranger*.

audience, which results in the use of a more formal language. The beginning of the excerpt of the hotline call illustrates another difference between the two subregisters: In a hotline call, the structure of the conversation typically follows a given script. After being welcomed by the host, the caller directly proceeds to ask a question on a highly specific medical topic. Neither the topic, nor the fact that the caller queries the host for information need to be negotiated or commented on by the speakers. While it is possible to imagine such a conversation between speakers that share an amicable relationship, such a query will likely be followed by a question as to why the caller asked the question in the first place.

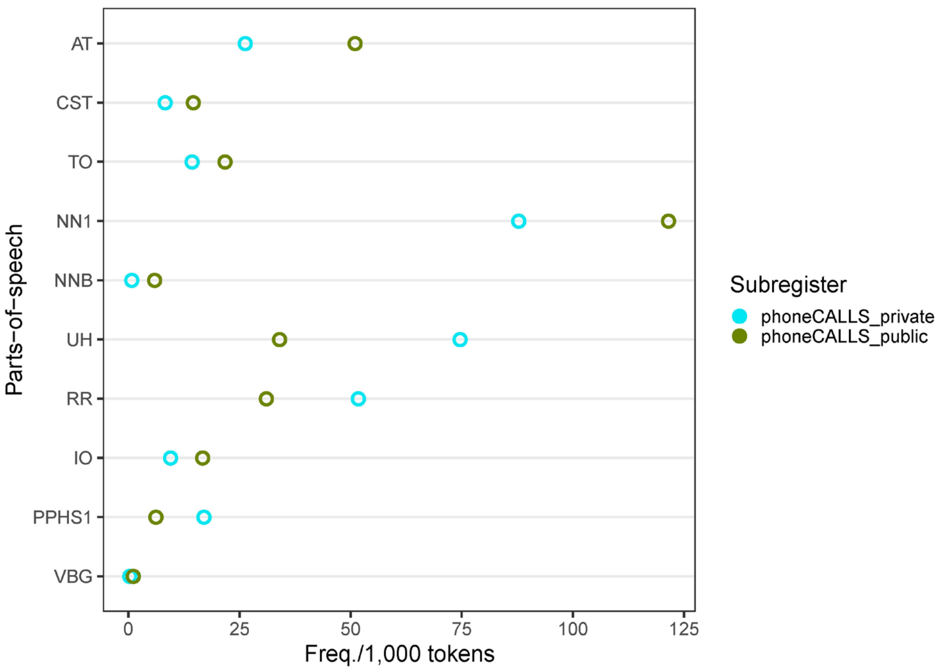


Figure 40: Top 10 POS tags that show the greatest differences between the subregisters. ICEtree function "Merge varieties (i.e. conflate corpora)" enabled.

Excerpt of a public phone call

<ICE-JA:S1A-100#1:1:A> Hello
<ICE-JA:S1A-100#2:1:B> Good evening doctor
<ICE-JA:S1A-100#3:1:A> Good evening
<ICE-JA:S1A-100#4:1:A> Welcome to our show
<ICE-JA:S1A-100#5:1:B> Uhm two questions
<ICE-JA:S1A-100#6:1:A> Sure go right ahead
<ICE-JA:S1A-100#7:1:B> Can a person's fingernail reflect one's state of health and if so what for example would be a warning sign
<ICE-JA:S1A-100#8:1:A> Haha
<ICE-JA:S1A-100#9:1:A> Uhm that's a very interesting question because in what's known as traditional Chinese <{1> <[1> medicine</[1> if you went to <}> <-> a</-> <=>

a</=></}> doctor in China<{2><[2><,></[2> his examination might involve taking a history that's asking about your complaints<{3><[3><,></[3> looking at your tongue and looking at your fingernails<{4><[4><,></[4> and feeling your pulse<{5><[5><,></[5> and just from doing those things he often is able to make a very accurate analysis of your state of health<{6><[6><,></[6>
<ICE-JA:S1A-100#10:1:A> Even in conventional medicine<,> examining the fingernails can tell a lot about things like the circulation about whether you are anaemic or not<{7><[7><,></[7> about whether your body might be lacking in certain minerals about what potential problems with your hormones
<ICE-JA:S1A-100#11:1:A> There's a tremendous amount of information that can come from examining the fingernails
<ICE-JA:S1A-100#12:1:A> <{8> <[8> Sure</[8>
<ICE-JA:S1A-100#13:1:B> <[1> Mhm</[1></{1>
<ICE-JA:S1A-100#14:1:B> <[2> Yes</[2></{2>
<ICE-JA:S1A-100#15:1:B> <[3> Yes</[3></{3>
<ICE-JA:S1A-100#16:1:B> <[4> Yes</[4></{4>
<ICE-JA:S1A-100#17:1:B> <[5> Yes</[5></{5>
<ICE-JA:S1A-100#18:1:B> <[6> Yes</[6></{6>
<ICE-JA:S1A-100#19:1:B> <[7> Yes</[7></{7>
<ICE-JA:S1A-100#20:1:B> <[8> <}> <-> You</-></[8></{8> <=> you</=></}> know actually tell-tale signs I mean looking at the nails I mean it's the lines on it uh and so on and so on
<ICE-JA:S1A-100#21:1:A> Sure sure
<ICE-JA:S1A-100#22:1:A> Many many tell-tale signs <}> <-> for <.> o</.></-> <=> for one</=></}> thing the colour of your nailbed<{1><[1><,></[1>
<ICE-JA:S1A-100#23:1:A> Okay
<ICE-JA:S1A-100#24:1:A> Brittleness of the nail whether the nails break easily<{2><[2><,></{2> ridges discoloration or marks on the nail<{3><[3><,></[3> are all the kinds of things that the doctor will be looking for<{4><[4><,></[4>

Excerpt of a private phone call

<ICE-CAN:S1A-099#1:1:B> What
<ICE-CAN:S1A-099#2:1:A> You can get in trouble for that
<ICE-CAN:S1A-099#3:1:B> That's right
<ICE-CAN:S1A-099#4:1:B> Anyways <,> when <@> Ryan </@> talked to my mom yesterday they had like pretty lengthy talk I guess and uh <,>
<ICE-CAN:S1A-099#5:1:B> Like my mom doesn't like <@> Ryan </@>
<ICE-CAN:S1A-099#6:1:A> Mm hmm

<ICE-CAN:S1A-099#7:1:B> But I think she's <}> <-> gonna </-> <+> going to </+> </}> start liking him <,> <{> <[> when he gets back </[>
<ICE-CAN:S1A-099#8:1:A> <[> Really </[> </{>
<ICE-CAN:S1A-099#9:1:B> Yeah cos I think that she <}> <-> kinda </-> <+> kind of </+> </}> hit a turning point yesterday <,> when she was talking to him
<ICE-CAN:S1A-099#10:1:B> Because they were talking and uh <,> my mom told him to make sure he took good care of me when he took me to Windsor for the weekend
<ICE-CAN:S1A-099#11:1:A> Good
<ICE-CAN:S1A-099#12:1:B> But she hadn't said anything to me about whether I could go or not
<ICE-CAN:S1A-099#13:1:A> <{> <[> Well obviously </[> it's yes
<ICE-CAN:S1A-099#14:1:B> <[> <?> She's obviously </?> </[> </{> come to the conclusion that I'm going anyway so <,>
<ICE-CAN:S1A-099#15:1:A> <{> <[> Yeah </[>
<ICE-CAN:S1A-099#16:1:B> <[> She also </[> </{> talked to him about like what he's <}> <-> gonna </-> <+> going to </+> </}> do when he gets out of the military and stuff and how <,> she really didn't want me to marry someone in the military because <,> like I didn't know what the military was like and she didn't really want me to find out what it was like and just all this <}> <-> kinda </-> <+> kind of </+> </}> stuff

5.3. Discussion

In the course of this chapter, I have illustrated that the ICE text categories *press editorials, skills and hobbies, business transactions* and *phone calls* contain distinct subregisters, that the composition of these text categories differs markedly between components of ICE and that these differences would be detectable if the corpus texts were annotated with situational characteristics. The analyses further showed that the connection between the concrete realization of texts and their situational characteristics is indeed a strong one. Except for a few outlier texts, the POS profiles cluster by subregister in the MDS solutions presented in the individual sections of this chapter. This pattern holds even for closely related text types such as *institutional editorials* and *personal editorials* or *private phone calls* and *public phone calls.* In the remainder of this section, I will briefly summarize the findings of the case studies and review the results from a broader perspective.

Concerning *press editorials,* where the ICE components ICE-GB, ICE-USA, ICE-HK and ICE-IND only include *institutional editorials,* ICE-CAN includes samples from a variety of opinion pieces (similar to the sampling scheme of the Brown corpus family), and ICE-JA only *personal editorials.* Compared to their institutional counterpart, personal opinion pieces are characterized by features connected to a more informal und personal style and address the reader more directly (especially 1st and 2nd person pronouns). These findings largely corroborate earlier studies and are in line with the conceptualizations of practitioners in the field (see section 5.2.1.1) as well as with studies that took different subregisters of press editorials into account (esp. Biber 1988, Morley & Murphy

2011, Sigley 2012). With respect to the corpus category *skills and hobbies*, a similar pattern was observable. Of the four ICE components investigated, only ICE-GB and ICE-JA exclusively contain instructional texts – even though the latter also includes cooking recipes. Approximately half the texts in ICE-CAN are purely informative and do not aim to directly instruct the reader. ICE-IND shows the broadest composition in terms of subregisters as it comprises instructional texts, cooking recipes as well as informative texts. For the register *business transactions*, ICE-IND largely contains reports that are read out by individuals in board meetings at educational institutions. The sampled texts in the other components investigated (ICE-GB, ICE-CAN & ICE-JA) are of a much more conversational character where individuals discuss business or work-related topics interactively in a group. The transcription practices of the Jamaican *business transactions* differ from the other components as interjections seem not to be included. Additionally, the Jamaican texts exhibit a strong topical focus (wage negotiations) and the speakers represent institutions. *Phone calls* was divided into two distinct subregisters: *private* and *public phone calls*. ICE-JA contains exclusively *public phone calls* – i.e. calls to hotlines or radio shows. ICE-GB, ICE-CAN and ICE-IND, on the other hand, comprise exclusively private conversations among colleagues, friends or family members.

What the subregister differences in the ICE sections analysed have in common is that the distance between the communication participants in one subregister in each of the discussed sections is smaller than for the other subregister(s). The notion of distance is, like register itself, an abstract and multi-faceted concept. Distance between communication participants can be spatial, temporal, situational, social or a combination of these and can be estimated from the situational characteristics of a text. A difference in distance is most obvious for the register *phone calls*: The interlocutors of *public phone calls* in ICE-JA very likely talk to each other for the first time and share very little if any personal knowledge. The opposite is the case for the subregister *private phone calls* in ICE-GB, ICE-CAN and ICE-IND. The distance between the communication partners is smaller in this subregister as the interlocutors share a personal relationship and personal knowledge. Such a difference in the situation of communication, which can be operationalized by the annotation of the situational characteristics of the texts, prompts the use of different linguistic structures. Although register is an inherently multivariate phenomenon, the distance between the communication partners appears to have a large influence on the realization of a text. This issue is also addressed by Koch & Oesterreicher (1985, 2012), who argue in a top-down fashion that the combination of certain conditions of communications (or situational characteristics) lead to certain strategies of verbalization (and thus the use of certain linguistic structures), and that register variation essentially occurs in “a *multi-dimensional* space between two poles” (Koch & Oesterreicher 2012: 447). They refer to these poles as *the language of immediacy* and *the language of distance* respectively. Figure 41 shows a modified version of the communicative model presented by Koch and Oesterreicher (2012: 450). Compared to the original model, the hypothesized conceptual space (represented by the rectangle in the middle) is slightly rotated to indicate that while there is overlap between the continuum of spoken and written registers, the empirical findings of the present study suggest that this overlap is not as large as suggested in the original visualization of Koch and Oesterreicher.

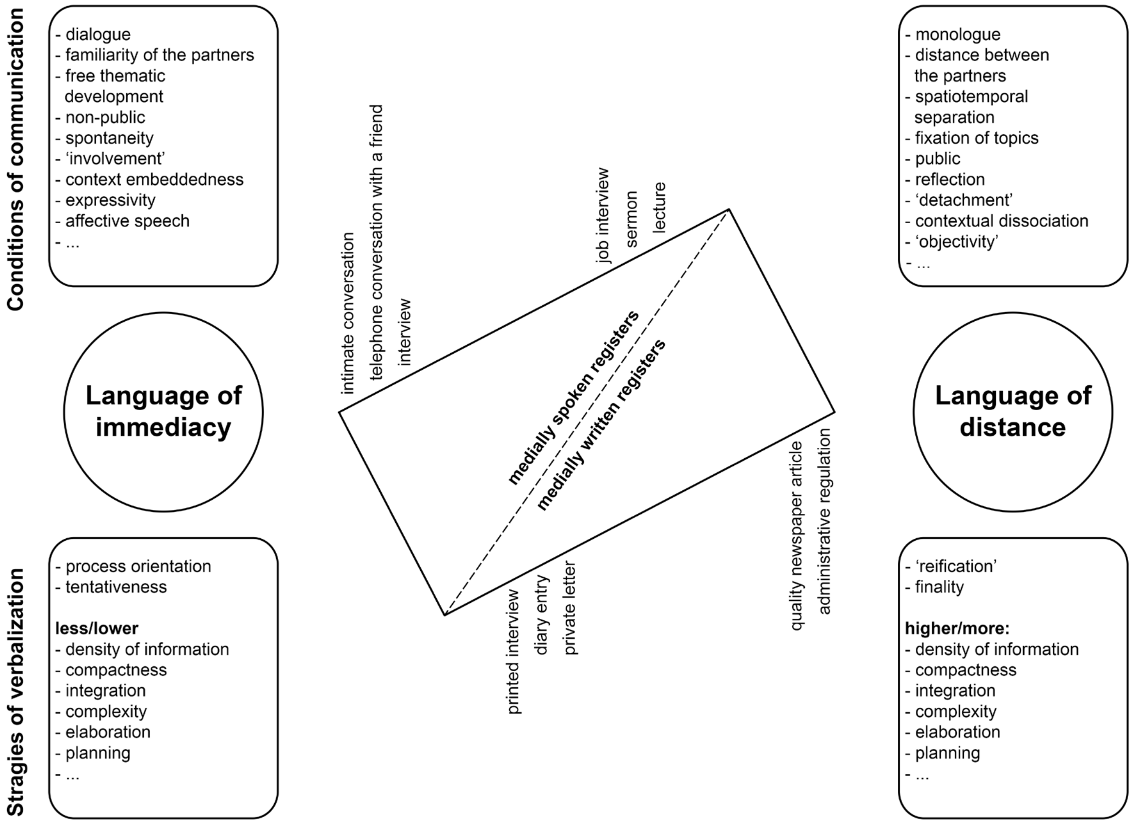


Figure 41: Slightly modified version of the communicative model (adapted from Koch and Oesterreicher 2012: 450).

In general terms, the language of immediacy is characterized by situations of communication of high immediacy – i.e. involved, dialogic, private and spoken communication between communication partners who share a high degree of familiarity – and prompt strategies of verbalization connected to involved language. In contrast, the language of distance is characterized by situations of communication which exhibit a high distance – i.e. monologic, highly explicit, asynchronous, public and written communication – and prompts strategies of verbalization connected to highly explicit, informational language production. Examples for both poles would be a private conversation among friends and a newspaper article or an administrative regulation respectively. Despite the intuitive association of medially written registers with the distance pole and medially spoken registers with the immediacy pole, Koch & Oesterreicher (2012) assert that, when positioned in this conceptual space, registers of both media may overlap to some extent as regards their linguistic layout and that some written registers (e.g. private letters) are connected more closely with the language of immediacy than some spoken registers (e.g. a university lecture) and vice versa.

Interestingly, this theorized overlap also surfaced in empirical bottom-up approaches to register variation: It is clearly present in the distribution of registers along Dimension 1 ("Involved vs.

Informational) of Biber's (1988) original MDA of the LOB and LLC corpus, in the results of the HCA of the registers of ICE-GB based on word *n*-grams in Gries et al. (2011), and also in the POS-based approach presented in the present study (see especially Figures 4 and 5 in section 3.2.1). These findings are significant in this context for three reasons: i) they support the notion that the immediacy or distance of a situation of communication is a major factor in the choice of verbalization strategies (and therefore the linguistic features) a speaker employs, ii) while register variation is often investigated using frequencies of more complex linguistic features (such as the occurrence of *that*-deletion, contractions, present tense verbs or private verbs),[37] it is also measurable with very simple metrics such as the frequencies of word or POS *n*-grams, and iii) it highlights the importance of a detailed documentation of the sampling process of corpus registers as even small variations in the situational characteristics of a situation of communication affects the concrete realization of texts.

When viewing the findings of the case studies presented in this section in the context of other registers, I would then expect to find that the subregisters characterized by a higher immediacy between the communication participants are closer to the register *conversation* in an MDS solution and the subregisters characterized by a higher distance between the communication participants closer to the ICE register *academic writing*. Figure 42 presents an MDS solution where each symbol indicates the aggregated POS profile of all texts in a subregister. The subregisters plotted in this figure include the texts analysed in this chapter plus the profiles of private conversations (*conversation*) and academic prose in the domain of natural sciences (*academicNATSCIENCE*) as instantiations of the two poles hypothesized by Koch & Oesterreicher (2012).

[37] This essentially applies to all linguistic features with a high loading on Dimension 1 in Biber's original MDA (cf. Biber 1988: 89) as this dimension closely mirrors the conceptual space hypothesized by Koch & Oesterreicher (2012).

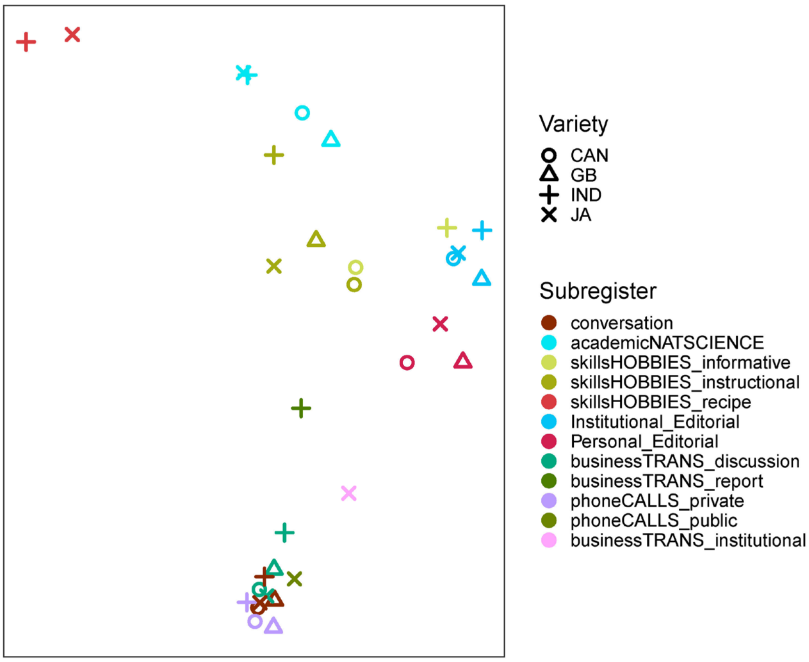


Figure 42: MDS plot of POS profiles of *private conversation*, *academic writing* (NATSCIENCE) and subregisters of *press editorials*, *skills and hobbies*, *business transactions* and *phone calls* of ICE-CAN, ICE-GB, ICE- JA and ICE-IND.

Although this study only investigated the variation *within* an ICE text category on the lowest level of the ICE sampling scheme, the expected tendencies are visible nevertheless. The interactive business discussions (*businessTRANS_discussion*) are closer to private conversations than institutional business discussions (*businessTRANS_institutional*), and the read-out reports (*businessTRANS_report*) are located more distantly from private conversations. In the subregisters *businessTRANS discussion* and *businessTRANS institutional* the interlocutors directly engage in face-to-face conversations. In *businessTRANS report* a single speaker is reading out a report and thus does not directly engage in a conversation with the other persons in the meeting. Although the participants of the meeting are in the same room and most likely share a similar relationship as the interlocutors in the two interactive subregisters, the roles of the participants are different in the report setting. Additionally, the production circumstances differ markedly between the interactive discussions and the reports: While the texts in all three subregisters are medially spoken, the reports were written beforehand and are read out whereas the discussions are the result of ad-hoc language production.

As for press editorials, the subregister *personal editorials* is located more closely to private conversations than *institutional editorials*. Although the conceptualizations of register presented in this study generally assert that language users adapt their language to the situational context and not the other way around, authors of personal editorials very likely make conscious use of specific strategies of verbalization connected to the language of immediacy (e.g. the use of personal pronouns, contractions and intensifiers) to reduce the distance to their readership. Koch & Oesterreicher (2012: 451) refer to this artificial reduction of distance between communication partners as '*manufactured* immediacy' (see also Werner 2021 on the conscious use of linguistic features in performed language to shape the situation of communication).

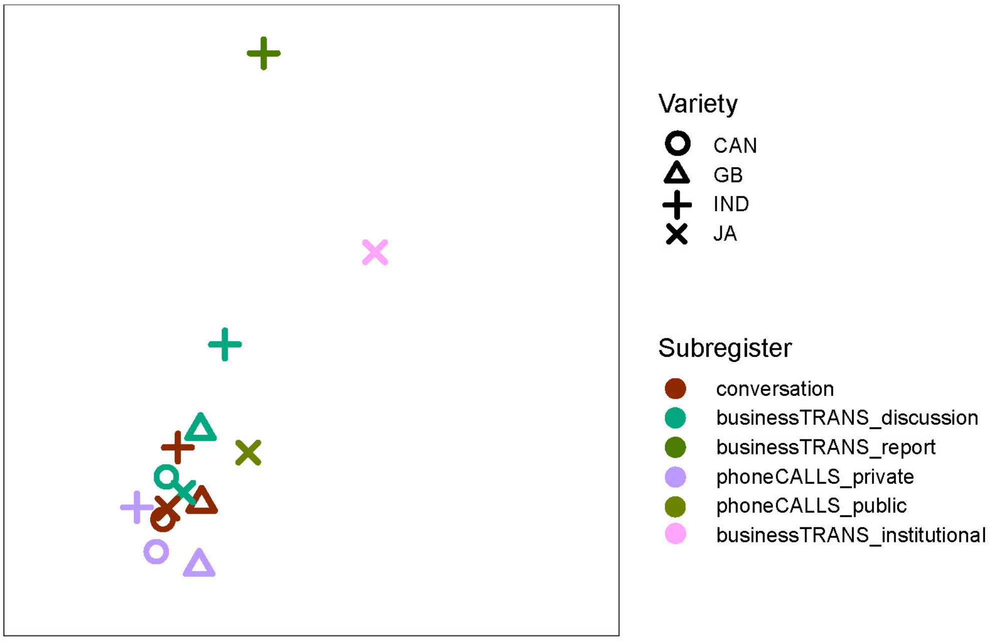


Figure 43: Lower half of Figure 42, zoomed in.

As the spoken subregisters in Figure 42 cluster densely, Figure 43 zooms in on the lower half of Figure 42. The register *phoneCALLS_public* is located slightly more towards the top of the plot and thus further away from the hypothesized immediacy pole. Interestingly, *phoneCALLS_private* is even further away from academic texts than *conversation*. Although this is again the point where detailed metadata of the situational characteristics of the texts are required to explain this pattern, the metadata of ICE-GB hint at a possible explanation: the register *conversation* not only contains face-to-face conversations among friends and family members, but also conversations between doctors and patients (e.g. ICE-GB:S1A-087-089) and counselling interviews (e.g. ICE-GB:S1A-059-060,-061), whereas the *phoneCALLS_private* in ICE-GB consist almost exclusively of conversations among friends and family (cf. Nelson et al. 2002: 310–312). The pattern then would fit well with the assumption that private conversations among individuals who share a close relationship represent a prototypical instantiation of involved language and thus the immediacy pole.

For the subregisters in *skills and hobbies*, Figure 42 shows a different pattern. Except for *skillsHOBBIES_recipe*, all other subregisters in Figure 42 scatter around an axis between *conversation* and the academic texts (*academicNATSCIENCE*). The subregister *skillsHOBBIES_recipe*, however, is located left of this imagined axis, and to a minor degree even further away from *conversation* than the assumed extreme of academic prose. Compared to other registers, cooking recipes represent a special case: Most commonly, cooking recipes consist of a set of detailed yet compact instructional statements. And although the situation of communication is similar to other registers marked by a high distance between communication partners, the communicative functions of recipes are comparatively restricted. For example, authors of recipes do not try to argue and convince readers what the best course of action is or convey large amounts of information. It is thus highly unlikely that we find linguistic features associated with these communicative functions (e.g. more complex syntactic structures or longer and more complex noun phrases).[38] The profiles of the two remaining subregisters (*skillsHOBBIES_instructional* and *skillsHOBBIES_informative*) are more similar and thus located more closely to the academic texts than to private conversations and both subregisters appear next to each other and roughly equidistant from both poles in the MDS. From a text-external perspective, it is difficult to argue that the distance between the readership and the author is smaller for one of the two subregisters. On the one hand, the authors of texts in *skillsHOBBIES_instructional* directly address and instruct the reader, but they do so in a highly explicit frame of reference and have a heavy informational focus: the author of ICE-GB:W2D-018, for example, instructs the readers on how to perform car maintenance on a specific car model while also providing detailed information on the design and function of various parts. On the other hand, the authors of some texts in *skillsHOBBIES_informative* give detailed recollections of events as if telling the story to someone with whom they share personal knowledge. For example, the author of ICE-CAN:W2E-017 describes his experience of a fly-fishing trip, including their private thoughts and emotions. In contrast to the instructional texts, however, the authors do not directly address their readership via personal pronouns or imperatives.

Overall, the results of this chapter are thus in line with the predictions of the communicative model of Koch & Oesterreicher (2012) and the results of Biber's (1988) MDA of the LOB and LLC corpus. When taking other corpus registers into account, it becomes evident that – with the exception of cooking recipes – the subregisters of a corpus category are closely related. While the POS profiles of the subregisters in a corpus category form distinct clusters, these clusters are located in close proximity, which suggests a high textual similarity. From a text-external perspective, the subregisters of a corpus category are also closely related as only a few situational characteristics differ between the subregisters.

[38] Indeed, it may be questioned whether cooking recipes and other types of manuals (such as assembly instructions and operation manuals of technical devices) represent a form of discourse at all. These registers are by design a one-way form of communication with only a singular point of reference – i.e. for use in the situation for which they were composed.

6. Conclusion and outlook

At the outset of this study, I criticized the common practice in English CL for a lack of transparency and documentation of the data that are included in a corpus and that this lack hinders the evaluation of corpus comparability – which in turn might lead to a misinterpretation of corpus findings. I argued that details on the sampling process are critical for the evaluation and interpretation of results in corpus-based investigations – especially if corpora or sections of such are compared. I singled out two areas in which corpus-based studies could greatly benefit from a more detailed documentation. The first concerns the minutiae of the sampling process of individual corpora and relates to the question of *how exactly* the data were selected and collected. The second concerns the demarcation of the strata. This is tied to the definition of the linguistic concepts of register and genre as these are used as the basis for the stratification of the sample. I further illustrated how, in the absence of a detailed documentation, quantitative methods provide auxiliary information which can help to evaluate the comparability of components of ICE. The methods are additionally made available in the form of the computer program *ICEtree*.

Concerning the first issue, I found that, while a linguistic corpus is an empirical sample that aims to be representative of a larger population, the relevant terminology established in sampling theory is often not used in CL. As a first step, I thus outlined the fundamental principles of sampling theory and reviewed the sampling processes of popular English reference corpora against this background. The reviewed reference corpora have in common that the target population is defined as the entirety of language produced in a given country in a given time period. Compilers of reference corpora seem to further agree that the sample needs to be stratified and that the notion of register or situation of use should form the basis for the stratification. However, the exact definitions and number of registers or strata vary between corpora. Equally diverse are the employed sample designs and the information on the sampling process that is covered in the documentation. While random sampling in its various forms might be the *conditio sine qua non* of empirical samples in some disciplines, it requires the construction of a sampling frame – and this is not possible for many registers. As a result, some registers need to be sampled using non-probability sample designs. This does not necessarily mean that these samples are less reliable and representative than samples that were gathered with random sampling techniques, it simply means that the representativeness cannot be statistically estimated.

Components of the Brown family and the BNC, for instance, largely employ random sampling techniques and are accompanied by a relatively detailed documentation. For other corpora such as the COCA (cf. Davies 2009: 161–164) or components of ICE, details on the sampling process remain largely undocumented. In the absence of information on the sampling process, it must be assumed that these corpora were compiled using non-probability sampling. While not problematic in itself, the sample design has an effect on the range of statistical tools that can be applied to the data. For instance, providing confidence intervals for non-probability samples is often advised against, yet it is common practice in corpus-based studies.

On the other hand, Leitner (1992: 43) rightly points out that "[r]andom sampling in Brown-Lancaster-Oslo-Bergen-Survey of English Usage has not been shown [...] to lead to better results than non-random sampling". Assessing the effect of the sample design is a difficult if not impossible endeavour. Even if a random and a non-random sample of the same population were available, the non-random sample might purely by chance be just as representative as the random sample. Or it might not. Non-probability samples, due to their very nature, may vary greatly and unexpectedly in terms of their representativeness. Yet regardless of the sample design, Woods et al. (1986: 55) suggest that, in the absence of information that indicates the opposite, we treat any sample as if it were a probability sample – and this reflects the current practice in corpus-based studies. While the issues related to the sample design are not easily resolvable, a transparent and detailed account of how the data were collected would allow individual researchers to better assess the implications of the sample design and its potential effects on a given study. A thorough documentation could additionally help explain otherwise puzzling or unexpected findings. Concerning the register-based stratification, I criticized that the common practice in CL is to rely on opaque register labels instead of on an operationalized definition of sampled registers. I argued that the use of these opaque labels might lead to different registers being hidden behind the same label in different corpora or components of a corpus family. I illustrated this issue for the registers *press editorials*, *skills and hobbies*, *business transactions* and *phone calls* in some components of ICE. Since previous studies on text clustering show that a bag-of-words or bag-of-features approach is sufficient to accurately cluster texts by register, I utilized the frequencies of part-of-speech monograms as a means for detecting small-scale register differences in these ICE registers. I calculated distances between texts based on these frequencies and visualized these distances mostly via MDS. Whenever distinct clusters emerged, I tried to reconstruct the situational characteristics of the texts to establish potential differences in register. The reconstruction was based on close reading and review of facsimiles of the original texts where available. In all of the analysed registers, a different composition in terms of sampled subregisters was attested for different components of ICE. Transcription practices and a dense topical focus additionally had an effect on the MDS solutions. By and large, however, the transcription practices and use of mark-up of components of ICE are very similar, which is likely due to the detailed transcription guidelines available for the ICE project. Regarding the type of frequency profile, it emerged that regional variation appears to have little effect on POS profiles. In contrast, Vetter (to appear) found that the MDA profiles, which are available in *ICEtree* as well, are also sensitive to regional variation. If we could establish which linguistic features are sensitive to which text-external factors, this would allow us to compile frequency profiles for specific tasks. Where POS profiles can be used to detect differences in register, other types of profiles could be used to detect or track other large-scale patterns of variation and change.

While I successfully exploited the bond between text types and registers to detect compositional differences between components of ICE, some points merit discussion. Although the evidence of the case studies suggests the contrary, it is possible that different registers spawn the same text types. In other words, only because the POS profiles form a single cluster, this need not necessarily mean that there is no difference in register. However, to be able to investigate this, a large number of texts from different registers that are annotated for their situational characteristics would be required. To my knowledge, such a database is not yet available. In any event, the case studies also showed that the composition of even fuzzily defined registers in components of ICE overlaps to a large degree. So, while I argued that these undocumented systematic differences between components of ICE lead to a decrease in the comparability of these registers and components, when zooming out further, i.e. when visualizing the differences between aggregated and not individual frequency profiles of registers (cf. Figure 17), it emerged that the observed differences have only a small distorting effect overall. In other words, while it cannot be ruled out that the differences in terms of sampled registers may have an effect on previous ICE-based studies, it is likely to a minor degree.

However, these differences could be rendered more visible by annotating texts with their situational characteristics during corpus compilation. Biber & Egbert (2018), for instance, employ a classification scheme based on selected situational characteristics for identifying individual registers in a random sample of the GloWbE corpus. Even though I envisage the annotation with the full set of situational characteristics as described in Biber & Conrad (2019: 40), I highlighted only the characteristics that differ most prominently between the subregisters in my own studies. It has to be established whether this set of situational characteristics is indeed sufficiently detailed and practical for the compilation and analysis of future (reference) corpora. This form of annotation would additionally open new avenues of research as the effects of individual situational characteristics could be taken into account. Instead of relying on a fixed hierarchical sampling scheme, texts could be grouped selectively by specific situational characteristics and this categorization could be tailored to the needs of individual linguistic enquiries.

Another major aspect of this study is the development of *ICEtree*, a web application that allows users not only to reproduce the quantitative parts of this study, but also to apply the methods described in section 3.2 and 4.1 to registers and ICE components not analysed here. While *ICEtree* in its current form has proven to be a useful tool, its functions and included data are relatively limited. Adding new methods or corpora, for example, is a time-consuming process and requires familiarity with the internal logic of the app. Concerning the frequency profiles, another possible avenue of research would be to investigate how other types of variables behave. Other types of profiles could be based on, for example, word frequencies or *n*-grams. I would expect profiles that are based on word frequencies to be more sensitive to a topical focus. With regard to the future development of *ICEtree,* an alternative path could be to integrate the functionality of *ICEtree* into existing and widely used open-source corpus tools such as *CQPweb*. The creation of the required frequency profiles then could be done on the fly. The data selection could be done manually or it could be based on a corpus query that the user performed. This way, the similarity of texts that include the query term could be directly examined. Additionally, as many users are already familiar with this interface, a wider user base could benefit from the functionality of the program as presented here without having to learn a new program.

In general, and despite the criticism brought forth in this study, the ICE corpus family has deservedly become an invaluable resource and has spawned a substantial body of research in the description of World Englishes – a fact that is also reflected in the recent review of the ICE project (Kirk & Nelson 2017). The goal of the review was to evaluate future prospects of the project and it is based on data from questionnaires, which were distributed among participants in the ICE project. Among other things, the review concludes that the ICE project continues and that future components 'comply [...] as far and as fully as possible with agreed norms, procedures and protocols of the project' (Kirk & Nelson 2017: 3–4). Based on the responses to what the main objectives should be for the future, it further concludes that more consideration should be given to homogenization, especially with regard to minimizing register differences between corpora (Kirk & Nelson 2017: 7). But achieving homogeneity in terms of sampled text varieties in the individual registers is certainly no trivial task. The question is indeed whether this is practically feasible at all. Providing a detailed account of the sample design and the included registers, however, definitely is.

Appendices

Metadata of newly sampled texts

Variety	File	Author	Title	Source	Register	Date
GB	W2E-011	Edward Pearce	Bridesmaids on the make	The Guardian	Personal Editorial	10.02.1993
GB	W2E-011	Dr. J. G. Howe	A time to live and a time to die for Tony Bland	The Guardian	Personal Editorial	10.02.1993
GB	W2E-012	Simon Hoggart	Simon Hoggart	The Observer	Personal Editorial	03.03.1991
GB	W2E-012	Chris Smith MP	All honour to the gay crusader	The Observer	Personal Editorial	13.01.1991
GB	W2E-013	William Shawcros s	The trouble with John Pilger	The Observer	Personal Editorial	17.03.1991
GB	W2E-013	Alan Watkins	The Tories are getting like the Labour of old	The Observer	Personal Editorial	17.03.1991
GB	W2E-014	Sue Arnold	How to engineer a new romance	The Observer	Personal Editorial	14.02.1991
GB	W2E-014	Richard Ingrams	Richard Ingrams	The Observer	Personal Editorial	14.02.1991
GB	W2E-015	Melanie Phillips	But in whose best interests?	The Guardian	Personal Editorial	05.02.1993
GB	W2E-015	Edward Pearce	Bring back the smokestack	The Guardian	Personal Editorial	03.02.1993
JA	W2E-011		A GREAT JAMAICAN	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-011		NOT GOLD — BUT SILVER	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-011		INTEREST RATE POLICY	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-011		Dr. AUBREY McFARLANE	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-011		WORTHWHILE GESTURE	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-011		Golf galore	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		A GREAT ACHIEVEMENT	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		A gentleman's dilemma	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		A HOME FOR HEROES	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		FUSS OVER CONDOMS	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		VIGILANTE JUSTICE	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		REGGAE SUCCESS	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-012		Back-seat driving	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-013		THE EVIL OF COCAINE	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-013		THE BOSNIAN CRISIS	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-013		SETTING PRIORITIES	The Gleaner	Institutional Editorial	Aug 92/Feb 93

JA	W2E-013	CONSTRUCTIO N COSTS	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-013	OVERVOTING	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	THE TRADE ACCOUNT	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	MINISTERIAL MUSIC	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	CRIME AND SECRECY	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	PINDLING IS DEFEATED	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	GOVERNMENT ACCOUNTS	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	A SHARED SORROW	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-014	That national dish	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	TRAGEDY OF SOMALIA	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	GOVERNMENT SPENDING	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	SCHOOL CHALLENGE	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	OMINOUS TREND	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	The little boozy bee...	The Gleaner	Institutional Editorial	Aug 92/Feb 93
JA	W2E-015	INTEREST RATES	The Gleaner	Institutional Editorial	Aug 92/Feb 93
CAN	W2E-011	Finally a budget	Winnipeg Free Press	Institutional Editorial	17.09.1990
CAN	W2E-011	Labor turning green	Winnipeg Free Press	Institutional Editorial	17.09.1990
CAN	W2E-011	Finally the war is over	Winnipeg Free Press	Institutional Editorial	17.09.1990
CAN	W2E-011	No policemen	Winnipeg Free Press	Institutional Editorial	01.02.1993
CAN	W2E-011	Aiming for Ottawa	Winnipeg Free Press	Institutional Editorial	01.02.1993
CAN	W2E-012	Staying on?	Winnipeg Free Press	Institutional Editorial	02.02.1993
CAN	W2E-012	Football fiasco	Winnipeg Free Press	Institutional Editorial	02.02.1993
CAN	W2E-012	Valor at Hong Kong	Winnipeg Free Press	Institutional Editorial	03.02.1993
CAN	W2E-012	Threatened culture	Winnipeg Free Press	Institutional Editorial	03.02.1993
CAN	W2E-013	Examine council activity	Medicine Hat News	Institutional Editorial	01.02.1993
CAN	W2E-013	Enact predator law now	Medicine Hat News	Institutional Editorial	01.02.1993
CAN	W2E-013	One way to end rumor	Medicine Hat News	Institutional Editorial	02.02.1993
CAN	W2E-013	Reverse discrimination	Medicine Hat News	Institutional Editorial	02.02.1993
CAN	W2E-013	Mayor Grimm's defence	Medicine Hat News	Institutional Editorial	03.02.1993
CAN	W2E-013	Reaching from the past	Medicine Hat News	Institutional Editorial	03.02.1993
CAN	W2E-014	Government neglect	Lethbridge Herald	Institutional Editorial	02.01.1993

CAN	W2E-014	Where are the mediators?	Lethbridge Herald	Institutional Editorial	02.01.1993
CAN	W2E-014	Tough balancing act	Lethbridge Herald	Institutional Editorial	01.02.1993
CAN	W2E-014	Citizen Lougheed	Lethbridge Herald	Institutional Editorial	03.02.1993
CAN	W2E-014	The school pinch	Lethbridge Herald	Institutional Editorial	03.02.1993
CAN	W2E-014	The 5% solution	Lethbridge Herald	Institutional Editorial	05.02.1993
CAN	W2E-014	Ethanol alternative	Lethbridge Herald	Institutional Editorial	05.02.1993
CAN	W2E-015	Words are not enough	The Gazette (Montreal Gazette)	Institutional Editorial	04.02.1986
CAN	W2E-015	Strikes against the public	The Gazette (Montreal Gazette)	Institutional Editorial	04.02.1986
CAN	W2E-015	No loss for Quebec	The Gazette (Montreal Gazette)	Institutional Editorial	04.02.1986
CAN	W2E-015	A new calling for Avon	The Gazette (Montreal Gazette)	Institutional Editorial	05.02.1986
CAN	W2E-015	Parking before praying	The Gazette (Montreal Gazette)	Institutional Editorial	05.02.1986
CAN	W2E-015	Chop Big O's roof	The Gazette (Montreal Gazette)	Institutional Editorial	06.02.1986
CAN	W2E-015	Machismo is not at issue	The Gazette (Montreal Gazette)	Institutional Editorial	06.02.1986

Tag sets

CLAWS7

The CLAWS7 tagset is available at http://ucrel.lancs.ac.uk/claws7tags.html.

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that), in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)

DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g. any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these, those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	Germanic genitive marker - (' or 's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number, neutral for number (two, three...)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction, neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)

NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh-degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)

RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g. about, in)
RPK	prep. adv., catenative (about in *be about to*)
RR	general adverb
RRQ	wh-general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be...)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)

VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A, b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

MAT

More details on the MAT tagset and the tagging process are described in Nini (2014).

AMP	Amplifiers
ANDC	Independent clause coordination
AWL	Word length
CAUS	Causative adverbial subordinators
CONC	Concessive adverbial subordinators
COND	Conditional adverbial subordinators
CONJ	Conjuncts
DEMO	Demonstratives
DEMP	Demonstrative pronouns
DPAR	Discourse particles
DWNT	Downtoners
EMPH	Emphatics
EX	Existential there
FPP1	First person pronouns
GER	Gerunds
HDG	Hedges
INPR	Indefinite pronouns
JJ	Attributive adjectives
NEMD	Necessity modals
NN	Total other nouns
NOMZ	Nominalizations
OSUB	Other adverbial subordinators
PHC	Phrasal coordination
PIN	Total prepositional phrases
PIT	Pronoun it
PLACE	Place adverbials
POMD	Possibility modals
PRED	Predicative adjectives
PRMD	Predictive modals
RB	Total adverbs
SPP2	Second person pronouns

SYNE	Synthetic negation
THAC	That adjective complements
THVC	That verb complements
TIME	Time adverbials
TO	Infinitives
TOBJ	That relative clauses on object position
TPP3	Third person pronouns
TSUB	That relative clauses on subject position
TTR	Type-token ratio
VBD	Past tense
VPRT	Present tense
XX0	Analytic negation
BEMA	Be as main verb
BYPA	By-passives
CONT	Contractions
PASS	Agentless passives
PASTP	Past participial clauses
PEAS	Perfect aspect
PIRE	Pied-piping relative clauses
PRESP	Present participial clauses
PRIV	private verbs
PROD	Pro-verb do
PUBV	Public verbs
SERE	Sentence relatives
SMP	Seem | appear
SPAU	Split auxiliaries
SPIN	Split infinitives
STPR	Stranded preposition
SUAV	Suasive verbs
THATD	Subordinator that deletion
WHCL	WH-clauses
WHOBJ	WH relative clauses on object position
WHQU	Direct WH-questions
WHSUB	WH relative clauses on subject position
WZPAST	Past participial WHIZ deletion relatives
WZPRES	Present participial WHIZ deletion relatives

ICEtree

The computer program *ICEtree* can be accessed digitally via the *Forschungsinformationssytem* (FIS) of the University of Bamberg or via the Open Science Framework repository located at https://osf.io/ztfsx/. The print version of this book includes *ICEtree* on an accompanying CD-ROM.

References

Aarts, Bas. 2006. Conceptions of categorization in the history of linguistics. *Language Sciences* 28(4). 361–385.

Aggarwal, Charu C. 2018. *Machine learning for text*. Cham: Springer.

Alonso Belmonte, Maria Isabel. 2007. Newspaper editorials and comment articles: a “cinderella” genre? *Revista Electrónica de Lingüística Aplicada* 6(Extra 1). 1–9.

Anthony, Laurence. 2018. *AntConc*. Tokyo, Japan: Waseda University.

Aston, Guy & Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA* (Edinburgh textbooks in empirical linguistics). Edinburgh: Edinburgh University Press.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1–16.

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bawarshi, Anis S. & Mary J. Reiff. 2010. *Genre: An introduction to history, theory, research, and pedagogy* (Reference guides to rhetoric and composition). West Lafayette: Parlor Press.

Bell, Allan. 1991. *The language of news media* (Language in society). Oxford: Blackwell.

Berber Sardinha, Tony & Marcia Veirano Pinto (eds.). 2014. *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (Studies in corpus linguistics 60). Amsterdam: Benjamins.

Bhatia, Vijay K. 2003. *Analysing genre: Language use in professional settings*. Harlow, Munich: Longman.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27(1). 3–44.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5(4). 257–269.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–38.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style,* 2nd edn. (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web*: A corpus-based taxonomy of web registers*. *Corpora* 10(1). 11–45.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487–517.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Biber, Douglas & Randi Reppen (eds.). 2015. *The Cambridge handbook of English corpus linguistics* (Cambridge handbooks in language and linguistics). Cambridge: Cambridge University Press.

Bonyadi, Alireza. 2011. Linguistic manifestations of modality in newspaper editorials. *International Journal of Linguistics* 3(1).

Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications,* 2nd edn. (Springer Series in Statistics). New York: Springer.

Borg, Ingwer, Patrick J. F. Groenen & Patrick Mair. 2013. *Applied multidimensional scaling* (Springer Briefs in Statistics). Heidelberg: Springer.

Brons-Albert, Ruth & Nicole Marx. 2010. *Empirisches Arbeiten in Linguistik und Sprachlehrforschung: Anleitung zu quantitativen Studien von der Planungsphase bis zum Forschungsbericht* (Narr Studienbücher). Tübingen: Narr.

Bryant, David & Vincent Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In Roderic Guigó & Dan Gusfield (eds.), *Algorithms in bioinformatics* (Lecture notes in computer science 2452), 375–391. Berlin, London: Springer.

Buchstaller, Isabelle & Ghada Khattab. 2013. Population samples. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics,* 74–95. Cambridge: Cambridge University Press.

Burnard, Lou. 2000. *Reference Guide for the British National Corpus (World Edition).* http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf (5 March, 2020).

Cavnar, W. & J. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the third Annual Symposium on Document Analysis and Information Retrieval,* 161–177. Information Science Research Institute, University of Nevada.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie & Jonathan McPherson. 2019. *shiny: Web application framework for R: 1.4.0.*

Clancy, Brian. 2010. Building a corpus to represent a variety of a language. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics* (Routledge Handbooks in Applied Linguistics), 80–92. London: Routledge.

Clear, Jeremy. 1992. Corpus sampling. In Gerhard Leitner (ed.), *New directions in English language corpora: Methodology, results, software developments* (Topics in English linguistics 9), 21–32. Berlin, New York: Mouton de Gruyter.

Conrad, Susan. 2015. Register variation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics,* 309–329. Cambridge: Cambridge University Press.

Cotter, Colleen. 2010. *News talk: Investigating the language of journalism.* Cambridge: Cambridge University Press.

Cox, Trevor F. & Michael A. A. Cox. 2001. *Multidimensional scaling,* 2nd edn. (Monographs on statistics and applied probability 88). London, New York: Chapman & Hall/CRC.

Crosthwaite, Peter. 2016. A longitudinal multidimensional analysis of EAP writing: Determining EAP course effectiveness. *Journal of English for Academic Purposes* 22. 166–178.

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8(4). 259–265.

Daniel, Johnnie. 2012. *Sampling essentials: Practical guidelines for making sampling choices.* Los Angeles: SAGE.

Davies, Mark. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2). 159–190.

Denoual, Etienne. 2006. A Method to quantify corpus similarity and its application to quantifying the degree of literality in a document. *International Journal of Technology and Human Interaction* 2(1). 51–66.

Diller, Hans-Jürgen. 2002. Genre vs. text type: Two typologies and their uses for the newspaper reader. In Andreas Fischer, Gunnel Tottie & Hans M. Lehmann (eds.), *Text types and corpora: Studies in honour of Udo Fries*, 17–28. Tubingen: Narr.

Dowle, Matt & Arun Srinivasan. 2019. *data.table: Extension of 'data.frame'.*

Eggins, Suzanne. 1994. *An introduction to systemic functional linguistics.* London: Pinter.

Everitt, Brian. 2011. *Cluster analysis,* 5th edn. (Wiley series in probability and statistics). Oxford: Wiley.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. Birmingham.

Expert Advisory Group on Language Engineering Standards. 1996. *EAGLES guidelines.* http://www.ilc.cnr.it/EAGLES96/home.html (5 March, 2020).

Fang, Alex C. 1996. AUTASYS: Grammatical tagging and cross-tagset mapping. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 110–124. Oxford: Clarendon Press.

Fang, Alex C. & Jing Cao. 2015. *Text genres and registers: The computation of linguistic features.* Heidelberg: Springer.

Fartousi, Hassan & Francisco P. Dumanig. 2012. Rhetoric of daily editorials: A review study of selected rhetorical analyses on daily editorials. *Advances in Asian Social Science* 2(1). 373–376.

Francis, Winthrop N. & Henry Kučera. 1979. *Brown corpus manual.* http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM (8 April, 2019).

Fuller, Wayne A. 2009. *Sampling statistics.* Hoboken, New Jersey: Wiley.

Gilquin, Gaëtanelle & Stefan T. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.

Greenbaum, Sidney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7(3). 315.

Greenbaum, Sidney. 1991. ICE*: The International Corpus of English. English Today* 7(04). 3.

Greenbaum, Sidney. 1996. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 3–13. Oxford: Clarendon Press.

Gries, Stefan T. 2006. Exploring variability within and between corpora*: Some methodological considerations. Corpora* 1(2). 109–151.

Gries, Stefan T. 2009. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds.), *Proceedings of Corpus Linguistics*. University of Liverpool.

Gries, Stefan T. 2011. Methodological and interdisciplinary stance in corpus linguistics. In Vander Viana, Sonia Zyngier & Geoff Barnbrook (eds.), *Perspectives on corpus linguistics* (Studies in corpus linguistics 48), 81–98. Amsterdam: Benjamins.

Gries, Stefan T. 2013. *Statistics for linguistics with R: A practical introduction,* 2nd edn. Berlin: Mouton de Gruyter.

Gries, Stefan T. 2015. The most under-used statistical method in corpus linguistics*: Multi-level (and mixed-effects) models. Corpora* 10(1). 95–125.

Gries, Stefan T. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics (IJCL)* 15(4). 520–548.

Gries, Stefan T., John Newman & Cyrus Shaoul. 2011. N-grams and the clustering of registers. *Empirical Language Research* 5(1).

Grieve, Jack. 2014. A multi-dimensional analysis of regional variation in American English. In Tony Berber Sardinha, Marcia Veirano Pinto & Douglas Biber (eds.), *Multi-dimensional analysis, 25 years on a tribute to Douglas Biber* (Studies in corpus linguistics 60). Amsterdam, Philadelphia: Benjamins.

Haan, Pieter de. 1992. The optimum corpus sample size? In Gerhard Leitner (ed.), *New directions in English language corpora: Methodology, results, software developments* (Topics in English linguistics 9), 3–20. Berlin, New York: Mouton de Gruyter.

Hansen, Beke. 2018. *Corpus linguistics and sociolinguistics: A study of variation and change in the modal systems of World Englishes* (Language and Computers Ser). Boston: Brill.

Hardie, Andrew. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.

Hundt, Marianne. 2015. World Englishes. In Douglas Biber & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics,* 381–400. Cambridge: Cambridge University Press.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), vol. 29.1, 154–167. Berlin: Mouton de Gruyter.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23(2). 254–267.

James, Gareth, Daniela Witten, Trevor Hastie & Robert Tibshirani. 2013. *An introduction to statistical learning* (Springer Texts in Statistics 103). Heidelberg: Springer.

Johansson, Stig. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* http://clu.uni.no/icame/manuals/LOB/INDEX.HTM#lob8http://clu.uni.no/icame/manuals/LOB/INDEX.HTM#lob8 (23 July, 2019).

Kanoksilapatham, Budsaba. 2007. Introduction to move analysis. In Douglas Biber, Ulla Connor & Thomas A. Upton (eds.), *Discourse on the Move: Using corpus analysis to describe discourse structure*, 35–54. Amsterdam, Philadelphia: Benjamins.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.

Kirk, John & Gerald Nelson. 2017. *Review of the ICE Project 2016/17.*

Kirk, John & Gerald Nelson. 2018. The International Corpus of English project: A progress report. *World Englishes* 37(4). 697–716.

Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe-Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36(1). 15–43.

Koch, Peter & Wulf Oesterreicher. 2012. Language of immediacy - Language of distance: Orality and literacy from the perspective of language theory and linguistic history. In Claudia Lange, Beatrix Weber & Göran Wolf (eds.), *Communicative spaces: Variation, contact, and change : papers in honour of Ursula Schaefer*, 441–473. Frankfurt: Lang.

Köhler, Reinhard. 2013. Statistical comparability: Methodological caveats. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum & Pascale Fung (eds.), *Building and using comparable corpora*, vol. 44, 77–91. Berlin, Heidelberg: Springer.

Kruger, Haidee & Adam Smith. 2018. Colloquialization versus densification in Australian English: A multidimensional analysis of the Australian Diachronic Hansard Corpus (ADHC). *Australian Journal of Linguistics* 38(3). 293–328.

Lee, David Y. W. 2001. Genres, registers, text types, domain and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology: A Refereed Journal for Second and Foreign Language Educators* 5(3). 37–72.

Lee, David Y. W. 2010. What corpora are available? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics* (Routledge Handbooks in Applied Linguistics), 107–121. London: Routledge.

Leitner, Gerhard. 1992. International Corpus of English*: Corpus design - problems and suggested solutions*. In Gerhard Leitner (ed.), *New directions in English language corpora: Methodology, results, software developments* (Topics in English linguistics 9), 33–64. Berlin, New York: Mouton de Gruyter.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis.* Amsterdam, Philadelphia: Benjamins.

Levy, Paul S. & Stanley Lemeshow. 2008. *Sampling of populations: Methods and applications,* 4th edn. (Wiley series in survey methodology). Hoboken, New Jersey: Wiley.

Liaw, Andy & Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2(3). 18–22.

Ljung, Magnus. 2000. Newspaper genres and newspaper English. In Friedrich Ungerer (ed.), *English media texts, past and present: Language and textual structure* (Pragmatics & beyond 80), 129–214. Amsterdam: Benjamins.

Lüdeling, Anke & Merja Kytö (eds.). 2008. *Corpus linguistics: An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1). Berlin: Mouton de Gruyter.

Lüdeling, Anke & Merja Kytö (eds.). 2009. *Corpus linguistics: An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2). Berlin: Mouton de Gruyter.

Mair, Christian. 2006. Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora. In Antoinette Renouf & Andrew Kehoe (eds.), *The changing face of corpus linguistics* (Language and computers, studies in practical linguistics 55), 355–376. Amsterdam, New York: Rodopi.

Martin, J. R. 2001. Language, register and genre. In Anne Burns & Caroline Coffin (eds.), *Analysing English in a global context: A reader* (Teaching English language worldwide). London: Routledge.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice* (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.

McEnery, Tony & Andrew Wilson. 2004. *Corpus linguistics: An introduction,* 2nd edn. (Edinburgh textbooks in empirical linguistics). Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book* (Routledge applied linguistics). London: Routledge.

McNair, Brian. 2009. I, Columnist. In Bob Franklin (ed.), *Pulling newspapers apart: Analysing print journalism*, 112–120. London: Routledge.

Meyer, Charles F. 2004a. Can you really study language variation in linguistic corpora? *American Speech* 79(4). 339–355.

Meyer, Charles F. 2004b. *English corpus linguistics: An introduction* (Studies in English language). Cambridge: Cambridge University Press.

Meyer, Charles F. 2009. In the Profession*: The "empirical tradition" in linguistics. Journal of English Linguistics* 37(2). 208–213.

Moisl, Hermann L. 2015. *Cluster analysis for corpus linguistics* (Quantitative linguistics 66). Berlin: Mouton de Gruyter.

Montoro, Rocío. 2018. The creative use of absences. *International Journal of Corpus Linguistics* 23(3). 279–310.

Morley, John & Amanda Murphy. 2011. The peroration revisited. In Vijay K. Bhatia & Maurizio Gotti (eds.), *Explorations in specialized genres*, 199–216. Bern: Lang.

Müller, Horst. 2011. *Journalistisches Arbeiten: Journalistische Grundlagen journalistische Arbeitstechniken journalistische Darstellungsformen* (Reihe mediengestützte Wissensvermittlung 5). Mittweida: Hochschulverlag Mittweida.

Nelson, Gerald. 1996. The design of the corpus. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 27–35. Oxford: Clarendon Press.

Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English* (Varieties of English around the World). Amsterdam: Benjamins.

Nelson, Mike. 2010. Building a written corpus: what are the basics? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics* (Routledge Handbooks in Applied Linguistics), 53–65. London: Routledge.

Nini, Andrea. 2014. *Multidimensional analysis tagger.*

Nini, Andrea. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis: Research methods and current issues*, 67–94. London: Bloomsbury.

O'Keeffe, Anne & Michael McCarthy (eds.). 2010. *The Routledge handbook of corpus linguistics* (Routledge Handbooks in Applied Linguistics). London: Routledge.

Ostler, Nicholas. 2008. Corpora of less studied languages. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), 457–483. Berlin: Mouton de Gruyter.

Pang, W. L. 2019. *DesktopDeployR: A framework for deploying self-contained R-based applications to the desktop*. https://github.com/wleepang/DesktopDeployR (5 March, 2020).

Petrenz, Philipp & Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics* 37(2). 385–393.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasinger, Sebastian M. 2008. *Quantitative research in linguistics: An introduction* (Research methods in linguistics). London: Continuum.

Rayson, Paul & Roger Garside. 2000. Comparing corpora using frequency profiling. *Association for Computational Linguistics*. 1–6.

Reeves, Ian & Richard Keeble. 2014. *The newspapers handbook,* 5th edn. (Media practice). London: Routledge.

Richardson, John. 2009. Readers' letters. In Bob Franklin (ed.), *Pulling newspapers apart: Analysing print journalism*, 58–69. London: Routledge.

Ripley, Brian. 2019. *tree: Classification and regression trees.*

RStudio Team. 2019. *RStudio: Integrated development environment for R*. Boston.

Saitou, Naruya & Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4(4). 406–425.

Sand, Andrea. 1999. *Linguistic variation in Jamaica: A corpus-based study of radio and newspaper usage* (Language in performance 20). Tübingen: Narr.

Santini, Marina. 2004. A shallow approach to syntactic feature extraction for genre classification. *UK Special Interest Group for Computational Linguistics.*

Särndal, Carl-Erik, Bengt Swensson & Jan H. Wretman. 2003. *Model assisted survey sampling* (Springer Series in Statistics). New York: Springer.

Schlüter, Julia & Fabian Vetter. 2020. An interactive visualization of Google Books Ngrams with R and Shiny: Exploring a(n) historical increase in onset strength in a(n) huge database. *Journal of Data Mining and Digital Humanities* 7(Special issue on Visualisations in Historical Linguistics). 1–25.

Schlüter, Norbert. 2006. How reliable are the results? Comparing corpus-based studies of the present perfect. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 135–148.

Sievert, Carson. 2020. *plotly for R*.

Sigley, Robert. 1997. Text categories and where you can stick them*: A crude formality index. International Journal of Corpus Linguistics* 2(2). 199–237.

Sigley, Robert. 2012. Assessing corpus comparability using a formality index: The case of the Brown/LOB clones. In Shunji Yamazaki, Robert Sigley & Toshio Saito (eds.), *Approaching language variation through corpora: A festschrift in honour of Toshio Saito* (Linguistic insights), 65–114. Bern: Lang.

Sinclair, John. 2005. Corpus and Text: Basic Principles. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, 1–16. Oxford: Oxbow Books.

Steen, Gerard. 1999. Genres of discourse and the definition of literature. *Discourse Processes* 28(2). 109–120.

Straßner, Erich. 2000. *Journalistische Texte* (Grundlagen der Medienkommunikation 10). Berlin: Mouton de Gruyter.

Swales, John. 1991. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tang, Xiaoyan & Jing Cao. 2015. Automatic genre classification via n-grams of part-of-speech tags. *Procedia - Social and Behavioral Sciences* 198. 474–478.

Taylor, John R. 1995. *Linguistic categorization: Prototypes in linguistic theory,* 2nd edn. Oxford: Clarendon Press.

Taylor, John R. 2011. Prototype theory. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning* (Handbooks of Linguistics and Communication Science 33), 643–664. Berlin, Boston: Mouton de Gruyter.

Thompson, Geoff. 2014. Intersubjectivity in newspaper editorials*: Construing the reader-in-the-text*. In Freek van de Velde, Lieselotte Brems & Lobke Ghesquière (eds.), *Intersubjectivity and intersubjectification in grammar and discourse: Theoretical and descriptive advances* (Benjamins current topics 65), 77–100. Amsterdam: Benjamins.

Thompson, Steven K. 2012. *Sampling,* 3rd edn. (Wiley series in probability and statistics). Hoboken, New Jersey: Wiley.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work* (Studies in corpus linguistics 6). Amsterdam, Philadelphia: Benjamins.

Tognini-Bonelli, Elena. 2010. Theoretical overview of the evolution of corpus linguistics. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics* (Routledge Handbooks in Applied Linguistics), 14–27. London: Routledge.

Vetter, Fabian. To appear. Comparing approaches to (sub-)register variation: The press editorials sections in the British, Canadian and Jamaican components of ICE. In Julia Schlüter & Ole Schützler (eds.), *Data and methods in corpus linguistics: Comparative approaches*. Cambridge: Cambridge University Press.

Wahl-Jorgensen, Karin. 2009. Op-ed pages. In Bob Franklin (ed.), *Pulling newspapers apart: Analysing print journalism*, 70–78. London: Routledge.

Wattam, Stephen M. 2015. *Technological advances in corpus sampling methodology*. Lancaster: Lancaster University PhD.

Weisser, Martin. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis*. Chichester, West Sussex: Wiley.

Werner, Valentin. 2014. *The present perfect in world Englishes: Charting unity and diversity* (Bamberger Beiträge zur Linguistik 5). Bamberg: University of Bamberg Press.

Werner, Valentin. 2016. Rise of the undead? BE-perfects in World Englishes. In Valentin Werner, Elena Seoane & Cristina Suárez-Gómez (eds.), *Re-assessing the Present Perfect: Corpus studies and beyond* (Topics in English linguistics 91), 259-294. Berlin, Boston: Mouton de Gruyter.

Werner, Valentin. 2021. Text-linguistic analysis of performed language: revisiting and re-modeling Koch and Oesterreicher. *Linguistics* 59(3). 541–575.

Westin, Ingrid. 2002. *Language change in English newspaper editorials* (Language and computers 44). Amsterdam: Rodopi.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English*: Exploring cross-constructional variation and change. Diachronica* 30(3). 382–419.

Wolk, Christoph & Bridgit Fastrich. 2019. *ShinyConc*.

Woods, Anthony, Paul Fletcher & Arthur Hughes. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

Xiao, Richard. 2008. Well-known and influential corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), 383–456. Berlin: Mouton de Gruyter.

Zhou, S. 2012. 'Advertorials'*: A genre-based analysis of an emerging hybridized genre. Discourse & Communication* 6(3). 323–346.

Fabian Vetter | Issues of corpus comparability and register variation in the International Corpus of English:Theories and computer applications